Online Appendix

Appendix A. Examples of DIF Item Treatment

To report how researchers treat DIF items in practice, papers published in five American Psychological Association journals were reviewed: *Health Psychology* (HA), *Psychological Assessment* (PA), *Journal of Applied Psychology* (JAP), *Journal of Family Psychology* (JFP), and *Journal of Personality and Social Psychology* (JPSP). DIF study papers were searched with the keyword "differential item functioning" in "Search inside this journal" on each journal's web site. We selected 27 papers that reported DIF results for empirical data sets in these five journals. We did not search with the keyword "measurement invariance," which is a more common term in the use of the linear confirmatory factor model. Our study focus is on DIF treatment when a portion of items are detected as DIF items and categorical responses are collected. However, it is common practice to test measurement invariance for all items when the linear confirmatory factor model is used, even though there are exceptions where *partial* measurement invariance is tested as in DIF analysis (e.g., Kim & Yoon, 2011, Reise, Widaman, & Pugh, 1993, for the exception). For example, a weak invariance model and a strong invariance model are compared in order to test whether item locations for all items differ between the two groups.

Table 1 represents our survey results including the group of interest for DIF study, DIF detection methods, % of DIF items in a test (i.e., [number of detected DIF items/total number of items]×100), and DIF item treatment. As shown in Table 1, dominant groups of interest were gender and ethnicity, and 74.1% of studies used IRT DIF detection methods. There are five distinct practices to deal with DIF items: (a) delete DIF items (30%); (b) no further action[1] (33%); (c) ignore DIF items[2] (26%); (d) calibrate items for each group

[1] We categorized papers as "no further action" papers when a DIF treatment was not mentioned. It is possible that the study's purpose was to detect DIF items so that no further action is required to be mentioned in the same paper.

[2] We categorized papers as "ignore DIF items" papers when the unidimensionality of a test was assumed in the presence of DIF. Some authors concluded this after they showed that DIF is negligible at the test level (e.g., Cooke et al., 2001; Orlando & Marshall, 2002) or DIF effect sizes are small (e.g., Freeman et al., 2012).

(i.e., no further group comparison) (7%); and (e) model DIF (4%). It is expected that DIF treatment practice may depend on the number of DIF items. The number of DIF items in the papers we surveyed ranges from 5% to 56% in “delete DIF items”; from 1.4% to 66% in “no further action”; from 8% to 43% in “ignore DIF items”; from 50% to 70% in “calibrate items for each group”; and from 20% to 50% in “model DIF.”

Two dominant practices of “delete DIF items” and “no further action” can be problematic in validating a test’s psychometric properties. When there is a large portion of DIF items in a test, deleting DIF items may result in lowering reliability and content validity.[3] When DIF treatment was not discussed well as in papers categorized in “no further action,” test users are left to deal with DIF items. Ignoring DIF items is expected to result in biased item parameter estimates and person scores of the unidimensional item response model. As presented in Bolt et al. (2004) and Smith and Reise (1998), the “calibrate items for each group” practice using a multigroup item response model is a recommended solution to report the psychometric properties of a test in the presence of a large numbers of DIF items. However, this practice does not aim for a group comparison, assuming all persons are on the same scale. In addition, as pointed out by Smith and Reise (1998, p. 1360), multigroup analysis has a problem when people have response patterns more consistent with the other group’s item parameters than with their own group. As an example of “model DIF” practice, Nye and Drasgrow (2011) showed how total observed mean differences can be decomposed into DIF and impact in the use of confirmatory factor model. The DIF treatment presented in this paper can be considered a category of “model DIF” because the group difference and individual difference are estimated in the same model. However, Nye and Drasgrow (2011) did not model a secondary dimension separate from a primary dimension such that the group difference and individual scores in their model may not be meaningful for group comparisons.

[3] We view content validity as part of the construct validation process.

Table 1: Examples of DIF Item Treatment

Journal	Study	Group of Interest	Detection Method	% of DIF	DIF Treatment
HP	DeWalt et al. (2013)	gender	IRTLR	10%	delete DIF item
PA	Cooke et al (2001)	ethnicity	Logistic Reg.	25%	ignore DIF item;DTF analysis
PA	Orlando and Marshall (2002)	Spanish speakers vs. English speakers	IRTLR	35%	ignore DIF item;DTF analysis
PA	Mungas et al (2004)	education;ethnicity;gender;age	Logistic Reg.	43%	ignore DIF item;DTF analysis
PA	Bolt et al. (2004)	gender;forensic psychiatric vs. criminal offenders	IRTLR	50-70%	multiple-group IRT analysis
PA	Neal et al (2006)	gender;time points	IRTLR	22%	delete DIF item
PA	McCarthy et al. (2009)	grade;gender;ethnicity	IRTLR	5%	delete DIF item
PA	Wiesner et al. (2010)	ethnicity	M-CFA	17%	ignore DIF item
PA	Chiesi et al. (2012)	gender;age	IRTLR	8%	ignore DIF item
PA	Fledderus et al. (2012)	age	LM	20%	ignore DIF item
PA	Wu et al. (2012)	gender;race/urban status	LRT	30%	no further action
PA	Freeman et al. (2012)	socioeconomic, racially and clinically distinct samples	IRTLR	40%	ignore DIF item;small effect size
JAP	Ellis (1989)	ethnicity	Lord	1.4-7.6%	no further action
JAP	Ellis and Kimmel (1992)	single-culture vs. multicultural group	Lord	4%	no further action
JAP	Whitney and Schmitt (1997)	ethnicity	Logliner	27%	delete DIF item
JAP	Collins et al. (2000)	gender;ethnicity	Lord;DFIT	10%-60%	no further action
JAP	Donovan et al. (2000)	computerized vs. paper-pencil formats	DFIT	11%	delete DIF item
JAP	Facteau and Craig (2001)	rater group	M-CFA;DFIT	8%	no further action
JAP	Stark et al. (2001)	applicants vs. nonapplicants	Lord,SIBTEST,DFIT	22%	delete DIF item
JAP	Raju et al. (2002)	ethnicity	M-CFA;DFIT	10%-20%	no further action
JAP	Stark et al. (2004)	job applicants vs. nonapplicants;ethnicity	DFIT	15%-39%	no further action
JAP	Meade (2010)	country;administration format (paper vs. internet)	IRTLR	66%;0%	no further action
JAP	Nye and Drasgrow (2011)	country	M-CFA	20-50%	model DIF
JAP	Meade and Wright (2012)	time points	IRTLRDIF	10%	no further action
JFP	Bingenheimer et al. (2005)	ethnicity;gender;age	DFIT	13%;56%	delete DIF items
JPSP	Smith and Reise (1998)	gender	z-test	65%	multiple-group IRT analysis*
JPSP	Church et al. (2011)	country	M-CFA	40%-50%	delete DIF items

Note. *: delete DIF items for sensitivity checking, but discuss the possibility of separate DIF calibration

Appendix B. DIF Item Treatments under Current Practice

In the following specification, it is assumed that there is I_b number of DIF items among I number of items (i.e., $i = 1, \ldots, I_b$ for non-DIF items and $i = I_{b+1}, \ldots, I$ for DIF items). In addition, in the specification of multigroup and modeling approaches, it is assumed that a reference group is known and justified by researchers.

Deleting DIF Items

When DIF items are deleted in a test, the 2-parameter item response model is used only for non-DIF items $i = 1, \ldots, I_b$, specified as follows:

$$\text{logit}[P(y_{ji} = 1|\theta_j)] = \alpha_i \theta_j - \beta_i. \tag{1}$$

The latent variable, θ_j $(j = 1, \ldots, J)$, is often assumed to follow a normal distribution, and the mean and variance are set to 0 and 1, respectively, to identify the model.

Ignoring DIF Items

When DIF items are ignored (i.e., all DIF items are included for calibration), the 2-parameter item response model specified in Equation 1 is used for *all* items $(i = 1, \ldots, I)$.

Multigroup Analysis

For a reference group $(g = 1)$ and a focal group $(g = 2)$, a multigroup item response model (Bock & Zimowski, 1997) is specified as follows:

$$\text{logit}[P(y_{jig} = 1|\theta_{jg})] = \alpha_{ig} \theta_{jg} - \beta_{ig}. \tag{2}$$

The equality constraint on item parameters is imposed for non-DIF items $(i = 1, \ldots, I_b)$, whereas group-specific item parameters, α_{ig} and β_{ig}, are estimated for DIF items $(i = I_{b+1}, \ldots, I)$. The mean and variance of θ_{j1} for the reference group $(g = 1)$ are set to 0 and 1 in the normal distribution, respectively, for the model identification $(\theta_{j1} \sim N(0, 1))$. The mean and variance of θ_{j2} for the focal group $(g = 2)$ can be estimated with the constraints on the item parameters $(\theta_{j2} \sim N(\mu, \sigma^2))$.

Because of the non-DIF items, the person scores are on the same metric (i.e., comparable), and the estimated mean on the θ_{j2} (μ) is the impact between the reference group and the focal group. However, unlike the impact in the DIF modeling approach, the impact estimate in the multigroup analysis is on the “contaminated” dimension due to DIF. To make the impact meaningful, one can take a two-step approach. In the first step, item parameters for the reference group are estimated in the multigroup analysis. In the second step, person scores for all persons and the impact are obtained using the item parameter estimates for the reference group.

As mentioned earlier, in the multigroup analysis with the one-step approach, the mean and variance of θ_{j1} for the reference group are set to 0 and 1, respectively, in the normal (N) distribution ($\theta_{ji} \sim N(0, 1)$) to identify the models. With this constraint, the impact (μ) is on the standardized latent scale score. In the multigroup analysis with the two-step approach, however, this model identification constraint is not required because the item parameter estimates are assumed to be known in the second step. That is, the means and variances of θ_{j1} and θ_{j2} can be estimated ($\theta_{j1} \sim N(\mu_1, \sigma_1^2)$; $\theta_{j2} \sim N(\mu_2, \sigma_2^2)$). However, the mean and variance of θ_{j1} for the reference group are set to 0 and 1 for the scale comparability purpose with the impact from the modeling DIF approach.

Appendix C. DIF Items and Multidimensionality

To explain the multidimensionality that resulted from DIF for a secondary dimension modeling, we use an explanatory item response model for DIF analysis and its configuration (Meulders & Xie, 2004). Two distinct groups are assumed in the description of the model, and they are denoted as the *reference* and *focal* groups (Angoff, 1993, p. 11). The focal group refers to the particular group of interest, whereas the reference group refers to the group with whom the focal group is to be compared as a base group.

A 2-parameter item response model (without DIF items) can be specified as follows:

$$\text{logit}[P(y_{ji} = 1|\theta_j)] = \sum_{k=1}^{K} \alpha_k X_{ik} \cdot \theta_j - \sum_{k=1}^{K} \beta_k X_{ik}, \tag{3}$$

where k is an index for an item indicator ($k = 1, \ldots, K$), i is an index for an item ($i = 1, \ldots, I$), j is an index for a person ($j = 1, \ldots, J$), X_{ik} is an item indicator that equal 1 if $i = k$ and 0 otherwise, θ_j is a continuous latent variable (e.g., ability), α_k is an item discrimination parameter associated with X_{ik}, and β_k is an item location parameter associated with X_{ik}.

Person groups (reference group [R] and focal group [F]) and item groups (DIF items and non-DIF items) were created to formulate DIF:

- Person groups: reference group ($Z_j = 0$ for $j = 1, \cdots, J_a$) and focal group ($Z_j = 1$ for $j = J_a + 1, \cdots, J$)
- Item groups: non-DIF items ($X_{ik} = 0$ for $i = 1, \cdots, I_b$) and DIF items ($X_{ik} = 1$ for $i = I_b + 1, \cdots, I$)

For non-DIF items, the model is the same as Equation (3). The DIF effect was formulated as the coefficient of a person-by-item predictor, which is derived as the product of an item indicator (X_{ik}) and a person indicator (Z_j) indicating group membership. The model for items suspected of DIF ($i = I_a + 1, \cdots, I$) can be formulated as follows:

$$\text{logit}[P(y_{ji} = 1|\theta_j)] = (\sum_{k=1}^{K} \alpha_k X_{ik} + \delta_k^{(\alpha)} W_{jki}) \cdot \theta_j - (\sum_{k=1}^{K} \beta_k X_{ik} + \delta_k^{(\beta)} W_{jki}), \tag{4}$$

where W_{jik} is the product of a binary group indicator Z_j and an item indicator X_{ik} so that $W_{jki} = X_{ik} Z_j$, and $\delta_k^{(\beta)}$ and $\delta_k^{(\alpha)}$ are DIF effects for item location and discrimination, respec-

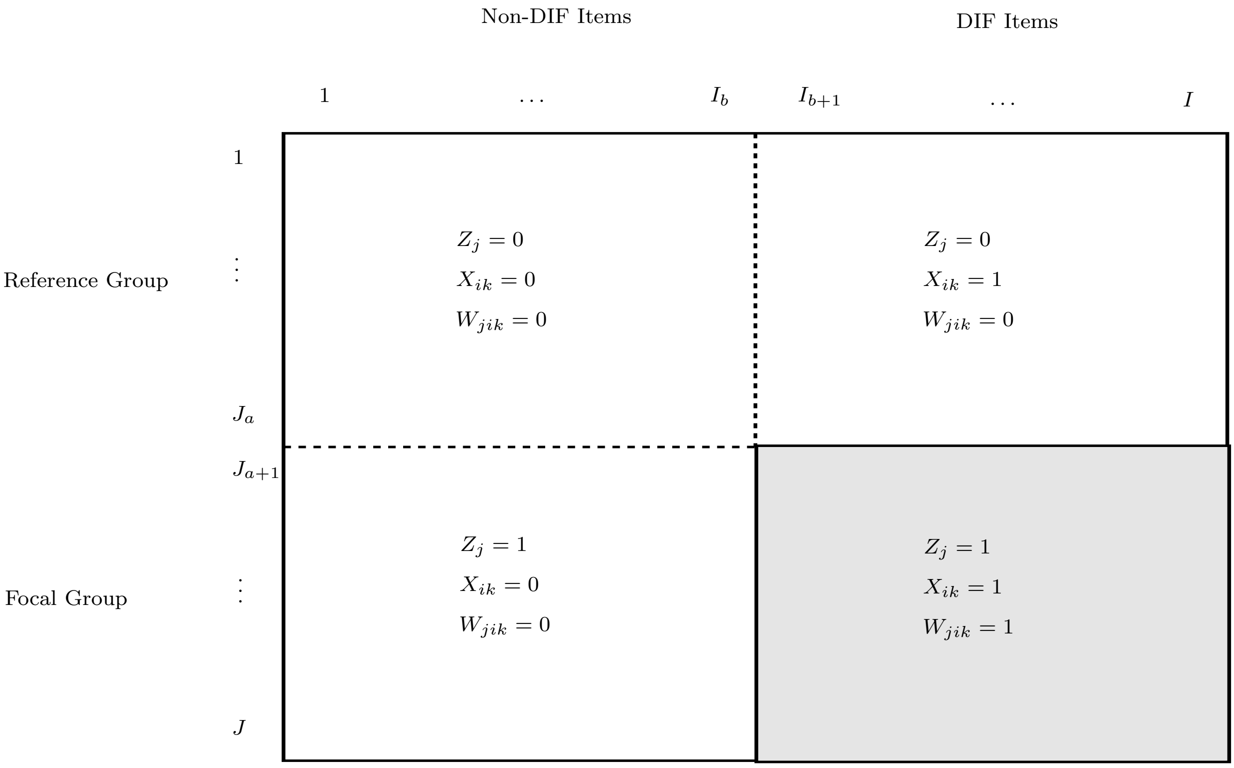


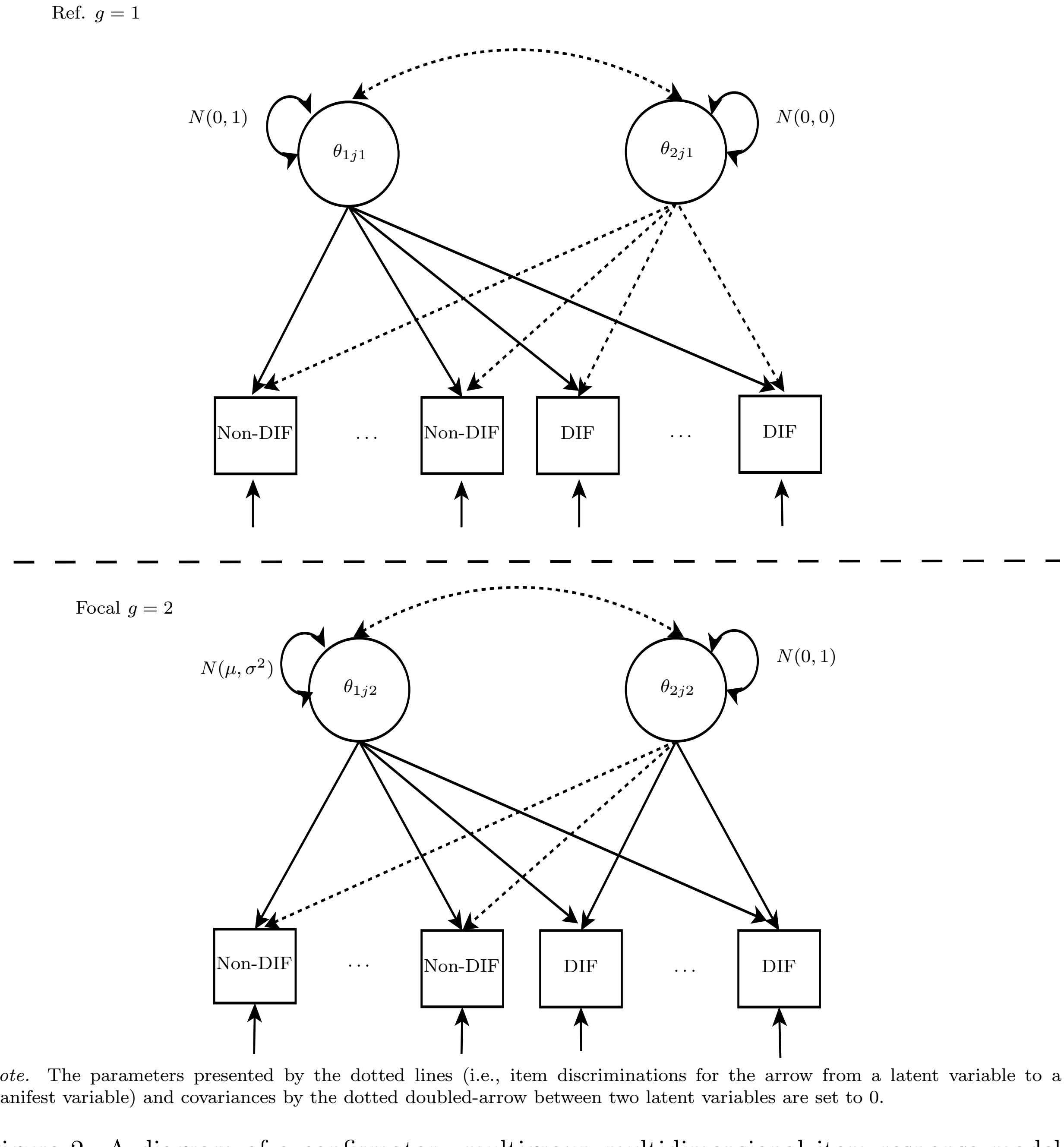
Figure 1: DIF configuration

tively. When DIF involves the discrimination ($\delta^{(\alpha)} \neq 0$) and possibly also item location, it is labeled *non-uniform* DIF, and when it involves only item location ($\delta^{(\beta)} \neq 0, \delta^{(\alpha)} = 0$)), it is labeled *uniform* DIF.

Figure 1 depicts the configuration of the IRT DIF models presented in Equations 3 and 4. As can be seen in Figure 1 (also Table 7.1. in Meulders & Xie, 2004), the logits of the endorsement probabilities are the same for three cells that have $W_{jki} = 0$, not for the right bottom cell that has $W_{jki} = 1$. Specifically, for the focal group, there are shifts with the DIF magnitudes on item parameters ($\delta_k^{(\alpha)}$ and $\delta_k^{(\beta)}$) for the items suspected of DIF. Additional dimension(s) other than θ_j can be considered to explain differences in endorsement probabilities for the focal group and DIF items, indicating multidimensionality exists as a result of DIF as a whole test.

Appendix D. A Diagram of a Confirmatory Multigroup Multidimensional Item Response Model

A separate measurement model is presented for the reference group ($g = 1$) and for the focal group ($g = 2$). In the figure, the squares and ellipses represent manifest and latent variables, respectively. The parameters presented by the dotted lines (i.e., item discriminations for the arrow from a latent variable to a manifest variable) and covariances by the dotted doubled-arrow between two latent variables are set to 0. Item responses represented by the squares are from the set of non-DIF items and the set of DIF items. In the measurement model of the reference group, dependency in item responses from all items is explained by θ_{1j1} (because all items have dashed lines for θ_{2j1}). In the measurement model of the focal group, dependency in the item responses from all items is explained by θ_{1j2} and dependency in item response from DIF items is explained by θ_{2j2} (because all non-DIF items have dashed lines for θ_{2j2}).



Note. The parameters presented by the dotted lines (i.e., item discriminations for the arrow from a latent variable to a manifest variable) and covariances by the dotted doubled-arrow between two latent variables are set to 0.

Figure 2: A diagram of a confirmatory multigroup multidimensional item response model for modeling DIF approach

Table 2: Appendix E. Comparisons among Four DIF Treatment Practices

DIF Practices		Parameter	Advantages	Disadvantages
Deleting		α,β	-	Calibrated for non-DIF items only
		μ	-	Not available
		θ	-	Lower reliability, lower content validity
Ignoring		α,β	Accuracy of estimates can be good with ignorable DIF effects.	Biased item parameter estimates due to DIF items
		μ	-	Not available
		θ	Accuracy of scores can be good with ignorable DIF effects.	Bias person scores due to DIF items
Multigroup	One-Step	α,β	Accuracy of estimates can be good within a group.	Only from the reference group→SE can be larger.
		μ	-	Not meaningful due to DIF items
		θ	Accuracy of estimates can be good within a group.	Not comparable between the two groups due to DIF items
	Two-Step	α,β	From both reference and focal groups	Required additional step
		μ	Meaningful comparison	Ignored uncertainty of item parameters
		θ	-	Required additional step
Modeling		α,β	From both reference and focal groups	Larger parameter variability due to model complexity
		μ	Meaningful comparison; Incorporated uncertainty of item parameters	-
		θ	Meaningful comparison; good content validity	-

“-” indicates that relevant information is not available.

Appendix E. An Illustrative Example

In the example, four practices for treating DIF, deleting, ignoring, multigroup analysis, and modeling DIF items, were compared. For the comparison of deleting DIF with other three practices, only non-DIF items were chosen in the other three practices. For all other comparisons that were not associated with the deleting DIF practice, all items were included for analysis. In comparing item parameter estimates across the four practices, item parameter estimates of the reference group were chosen in the multigroup analysis. For item parameter estimates and person score comparisons across the four practices, the root mean square difference (RMSD) for each pair of the four approaches was calculated for item parameter estimates and person scores.

Data and analysis. The data set was from a 72-item test called the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006) to measure face recognition ability across the entire range found in normal and abnormal populations. Cho et al. (2015) conducted Lord's chi-square test (Lord, 1980), Raju's z statistics (Raju, 1990), and the likelihood ratio test (LRT) method (Thissen, Steinberg, & Wainer, 1988) for age groups (younger [reference group]: age 20 years; older [focal group]: age $>$ 21 years). Twenty DIF items (28% of the CFMT test) were detected for the younger group (G1; reference group; $N = 1,271$) and the older group (G2; focal group; $N = 1,226$): 14 out of 20 items (Items 4, 5, 6, 7, 13, 21, 28, 29, 31, 35, 45, 49, 58, and 69) were detected as DIF items by any two of the three detection methods and 6 items (Items 22, 25, 32, 33, 57, and 64) were detected as DIF items by all three detection methods. There were six items with a large DIF magnitude (Items 5, 22, 32, 49, 54, and 57), based on the noncompensatory DIF (NCDIF) index proposed by Raju, van der Linden, and Fleer (1995).

Results. Table 3 presents item parameter estimates to illustrate the structure of item parameters for the modeling DIF approach. Table 4 shows the RMSD between the four practices for the item parameter estimates and person scores. For the RMSD with deleting practice, only 52 non-DIF items were considered as indicated by an asterisk in Table 4. The largest RMSD was found between ignoring DIF and multigroup analysis for item parameter estimates and person scores. The smallest RMSD was found between multigroup analysis and modeling DIF for item parameter estimates and person scores.

As noted earlier, impact cannot be obtained in deleting and ignoring DIF practices, whereas it can be estimated in multigroup analysis and modeling DIF. In the multigroup analysis, the impact was 2.095 (SE = 0.238, p-value = 0.000), which indicates that the mean of the older group was 2.095 higher than that of the younger group. However, this mean difference is not meaningful because they are not on the same construct because of DIF items. To overcome this limitation, the two-step approach can be used to estimate the meaningful impact in the use of multigroup analysis. The impact from the two-step approach was 0.276 (SE = 0.036, p-value = 0.000), which indicates that the mean of the older group was 0.277 higher than that of the younger group on the same (primary) construct. The method for modeling DIF yields 0.300 (SE = 0.051, p-value = 0.000), which means that the mean of the older group was 0.300 higher than that of the younger group on the same (primary) construct.

IRT reliability was calculated. The reliability of the deleting approach (0.854) was slightly lower than that of the ignoring approach (0.885) and the modeling approach (0.889). The reliability of the multigroup approach (with the two-step approach) was 0.834, which was lower than the reliability of the modeling approach.

Table 3: Example: Item Parameter Estimates (Standard Errors) of Modeling DIF Approach

		Reference			Focal		
	DIF?	α_1	α_2	β	α_1	α_2	β
1	No	0.97(0.66)	0.00	-5.49(0.54)	0.97(0.66)	0.00	-5.49(0.54)
2	No	1.57(0.30)	0.00	-4.71(0.33)	1.57(0.30)	0.00	-4.71(0.33)
3	No	1.78(0.27)	0.00	-4.55(0.30)	1.78(0.27)	0.00	-4.55(0.30)
8	No	1.24(0.26)	0.00	-4.48(0.27)	1.24(0.26)	0.00	-4.48(0.27)
9	No	2.24(0.51)	0.00	-5.85(0.65)	2.24(0.51)	0.00	-5.85(0.65)
10	No	2.05(0.32)	0.00	-5.50(0.42)	2.05(0.32)	0.00	-5.50(0.42)
11	No	1.42(0.21)	0.00	-3.98(0.22)	1.42(0.21)	0.00	-3.98(0.22)
12	No	1.72(0.26)	0.00	-4.69(0.30)	1.72(0.26)	0.00	-4.69(0.30)
14	No	1.41(0.19)	0.00	-3.69(0.19)	1.41(0.19)	0.00	-3.69(0.19)
15	No	1.75(0.31)	0.00	-4.96(0.36)	1.75(0.31)	0.00	-4.96(0.36)
16	No	2.10(0.30)	0.00	-5.44(0.39)	2.10(0.30)	0.00	-5.44(0.39)
17	No	1.57(0.17)	0.00	-3.68(0.18)	1.57(0.17)	0.00	-3.68(0.18)
18	No	2.45(0.28)	0.00	-5.58(0.36)	2.45(0.28)	0.00	-5.58(0.36)
19	No	2.45(0.32)	0.00	-5.55(0.42)	2.45(0.32)	0.00	-5.55(0.42)
20	No	1.39(0.10)	0.00	-2.10(0.09)	1.39(0.10)	0.00	-2.10(0.09)
23	No	1.22(0.07)	0.00	-0.43(0.06)	1.22(0.07)	0.00	-0.43(0.06)
24	No	1.09(0.08)	0.00	-1.50(0.07)	1.09(0.08)	0.00	-1.50(0.07)
26	No	1.77(0.17)	0.00	-3.54(0.17)	1.77(0.17)	0.00	-3.54(0.17)
27	No	1.40(0.10)	0.00	-2.17(0.10)	1.40(0.10)	0.00	-2.17(0.10)
30	No	1.28(0.09)	0.00	-1.60(0.08)	1.28(0.09)	0.00	-1.60(0.08)
34	No	0.43(0.04)	0.00	0.22(0.04)	0.43(0.04)	0.00	0.22(0.04)
36	No	0.98(0.06)	0.00	0.00(0.05)	0.98(0.06)	0.00	0.00(0.05)
37	No	0.48(0.05)	0.00	0.58(0.05)	0.48(0.05)	0.00	0.58(0.05)
38	No	1.31(0.11)	0.00	-2.29(0.10)	1.31(0.11)	0.00	-2.29(0.10)
39	No	0.96(0.06)	0.00	0.27(0.05)	0.96(0.06)	0.00	0.27(0.05)
40	No	1.59(0.12)	0.00	-2.59(0.11)	1.59(0.12)	0.00	-2.59(0.11)
41	No	1.35(0.09)	0.00	-1.83(0.08)	1.35(0.09)	0.00	-1.83(0.08)
42	No	1.57(0.08)	0.00	-0.15(0.07)	1.57(0.08)	0.00	-0.15(0.07)
43	No	1.00(0.06)	0.00	-0.69(0.06)	1.00(0.06)	0.00	-0.69(0.06)
44	No	1.60(0.10)	0.00	-1.23(0.08)	1.60(0.10)	0.00	-1.23(0.08)
46	No	0.84(0.05)	0.00	-0.09(0.05)	0.84(0.05)	0.00	-0.09(0.05)
47	No	0.93(0.07)	0.00	-1.32(0.06)	0.93(0.07)	0.00	-1.32(0.06)
48	No	1.61(0.09)	0.00	-1.16(0.08)	1.61(0.09)	0.00	-1.16(0.08)
50	No	0.77(0.05)	0.00	-0.44(0.05)	0.77(0.05)	0.00	-0.44(0.05)
51	No	0.46(0.05)	0.00	0.69(0.05)	0.46(0.05)	0.00	0.69(0.05)
52	No	1.13(0.08)	0.00	-1.12(0.07)	1.13(0.08)	0.00	-1.12(0.07)
53	No	0.90(0.06)	0.00	-0.99(0.06)	0.90(0.06)	0.00	-0.99(0.06)
54	No	1.03(0.07)	0.00	-0.86(0.06)	1.03(0.07)	0.00	-0.86(0.06)
55	No	1.36(0.08)	0.00	-0.51(0.06)	1.36(0.08)	0.00	-0.51(0.06)
56	No	0.93(0.06)	0.00	0.23(0.05)	0.93(0.06)	0.00	0.23(0.05)
59	No	0.70(0.05)	0.00	0.07(0.05)	0.70(0.05)	0.00	0.07(0.05)
60	No	1.39(0.08)	0.00	-0.79(0.07)	1.39(0.08)	0.00	-0.79(0.07)
61	No	0.83(0.05)	0.00	0.08(0.05)	0.83(0.05)	0.00	0.08(0.05)
62	No	0.74(0.06)	0.00	-1.49(0.06)	0.74(0.06)	0.00	-1.49(0.06)
63	No	1.28(0.08)	0.00	-1.68(0.08)	1.28(0.08)	0.00	-1.68(0.08)
65	No	0.83(0.06)	0.00	-1.26(0.06)	0.83(0.06)	0.00	-1.26(0.06)
66	No	1.02(0.06)	0.00	-0.30(0.06)	1.02(0.06)	0.00	-0.30(0.06)
67	No	0.82(0.06)	0.00	-1.38(0.06)	0.82(0.06)	0.00	-1.38(0.06)
68	No	0.65(0.05)	0.00	0.71(0.05)	0.65(0.05)	0.00	0.71(0.05)
70	No	0.57(0.05)	0.00	-0.53(0.05)	0.57(0.05)	0.00	-0.53(0.05)
71	No	1.40(0.08)	0.00	-0.25(0.06)	1.40(0.08)	0.00	-0.25(0.06)
72	No	0.46(0.04)	0.00	-0.04(0.04)	0.46(0.04)	0.00	-0.04(0.04)
4	Yes	1.76(0.27)	0.00	-4.97(0.34)	1.76(0.27)	0.58(0.51)	-4.97(0.34)
5	Yes	1.35(0.20)	0.00	-3.95(0.25)	1.35(0.20)	0.45(0.72)	-3.95(0.25)
6	Yes	1.40(0.24)	0.00	-4.26(0.36)	1.40(0.24)	0.76(0.76)	-4.26(0.36)
7	Yes	1.19(0.45)	0.00	-5.42(0.47)	1.19(0.45)	0.75(0.41)	-5.42(0.47)
13	Yes	1.45(0.15)	0.00	-3.23(0.15)	1.45(0.15)	0.14(0.20)	-3.23(0.15)
21	Yes	1.28(0.13)	0.00	-2.35(0.18)	1.28(0.13)	1.06(0.45)	-2.35(0.18)
22	Yes	1.09(0.08)	0.00	-1.43(0.07)	1.09(0.08)	0.09(0.90)	-1.43(0.07)
25	Yes	1.05(0.07)	0.00	-1.47(0.07)	1.05(0.07)	0.05(0.84)	-1.47(0.07)
28	Yes	0.81(0.06)	0.00	-1.13(0.06)	0.81(0.06)	0.17(0.39)	-1.13(0.06)
29	Yes	1.17(0.08)	0.00	-1.44(0.07)	1.17(0.08)	0.29(0.23)	-1.44(0.07)
31	Yes	1.17(0.08)	0.00	-0.81(0.06)	1.17(0.08)	0.33(0.35)	-0.81(0.06)
32	Yes	0.75(0.08)	0.00	-0.99(0.08)	0.75(0.08)	0.48(0.58)	-0.99(0.08)
33	Yes	1.43(0.12)	0.00	-2.11(0.13)	1.43(0.12)	1.14(0.31)	-2.11(0.13)
35	Yes	0.97(0.08)	0.00	-1.82(0.08)	0.97(0.08)	0.39(0.22)	-1.82(0.08)
45	Yes	1.10(0.11)	0.00	-2.57(0.11)	1.10(0.11)	0.70(0.24)	-2.57(0.11)
49	Yes	0.56(0.06)	0.00	-1.08(0.05)	0.56(0.06)	0.29(0.29)	-1.08(0.05)
57	Yes	0.40(0.04)	0.00	0.39(0.04)	0.40(0.04)	0.05(0.27)	0.39(0.04)
58	Yes	0.79(0.08)	0.00	-1.55(0.08)	0.79(0.08)	0.73(0.25)	-1.55(0.08)
64	Yes	1.18(0.07)	0.00	-0.34(0.06)	1.18(0.07)	0.11(0.15)	-0.34(0.06)
69	Yes	0.45(0.05)	0.00	0.28(0.04)	0.45(0.05)	0.05(0.39)	0.28(0.04)

Table 4: Example: Average Root Mean Square Difference (RMSD) of Item Discrimination Estimates (top), Item Location Estimates (middle), and Person Scores (bottom)

	Deleting	Ignoring	Multigroup	Modeling
Deleting				
Ignoring	0.078*			
Multigroup	0.073*	0.131		
Modeling	0.072*	0.100	0.089	

	Deleting	Ignoring	Multigroup	Modeling
Deleting				
Ignoring	0.069*			
Multigroup	0.189*	0.227		
Modeling	0.152*	0.174	0.132	

	Deleting	Ignoring	Multigroup	Modeling
Deleting				
Ignoring	0.158			
Multigroup	0.189	0.289		
Modeling	0.186	0.168	0.112	

Appendix F. True Item Parameter Examples: 10%, Nonuniform, and High Magnitudes

Table 5: True Item Parameter Examples: 10%, Nonuniform, and High Magnitudes

	Reference Group		Focal Group		Difference	
Item	α_1	β_1	α_2	β_2	α	β
1	1.444	-0.673	1.444	-0.673	0	0
2	1.419	0.402	1.419	0.402	0	0
3	1.310	0.061	1.310	0.061	0	0
4	0.869	0.980	0.869	0.98	0	0
5	0.461	-0.412	0.461	-0.412	0	0
6	0.906	0.889	0.906	0.889	0	0
7	2.669	-2.332	2.669	-2.332	0	0
8	1.592	0.921	1.592	0.921	0	0
9	0.715	0.051	0.715	0.051	0	0
10	1.967	0.274	1.967	0.274	0	0
11	0.935	-0.204	0.935	-0.204	0	0
12	1.125	-2.635	1.125	-2.635	0	0
13	0.173	0.477	0.173	0.477	0	0
14	1.441	0.013	1.441	0.013	0	0
15	1.078	0.551	1.078	0.551	0	0
16	1.241	-0.529	1.241	-0.529	0	0
17	0.643	0.007	0.643	0.007	0	0
18	0.749	-0.762	0.749	-0.762	0	0
19	0.982	1.103	0.382	2.103	0.6	-1
20	1.625	-1.113	1.025	-0.113	0.6	-1

Appendix G. DIF Effect Size Measures

In the current study, we assume that a subset of items are detected as DIF items for the purification method for item calibration and scoring. DIF effect size measures can be an indicator for researchers to decide which DIF treatment can be chosen among deleting, ignoring, multigroup, and modeling DIF. Scale (or test)-level DIF effect size measures were chosen to quantify how much DIF exists in the manipulated DIF conditions of a simulation study and consequently to compare differential behaviors of the different DIF treatment practices in the various DIF conditions at the scale (or test) level.

Two scale-level effect size measures, signed test difference in the sample ($STDS$) and unsigned test difference in the sample $UTDS$ (Meade, 2010), were chosen, and they were calculated using VisualDF (Meade, 2010). These two measures are the sums of the signed differences across items ($i = 1, \ldots, I$), specified as follows:

$$STDS = \sum_{i=1}^{I} SIDS_i \tag{5}$$

and

$$UTDS = \sum_{i=1}^{I} UIDS_i. \tag{6}$$

The item-level measures, $SIDS_i$ and $UIDS_i$, are:

$$SIDS_i = \frac{\sum_{j=J_a+1}^{J} (ES_{(ji|\tilde{\theta_j},\hat{\xi}_{Fi})} - ES_{(ji|\tilde{\theta_j},\hat{\xi}_{Ri})})}{J - J_a} \tag{7}$$

and

$$UIDS_i = \frac{\sum_{j=J_a+1}^{J} |ES_{(ji|\tilde{\theta_j},\hat{\xi}_{Fi})} - ES_{(ji|\tilde{\theta_j},\hat{\xi}_{Ri})}|}{J - J_a}, \tag{8}$$

where $ES_{(ji|\tilde{\theta_j},\hat{\xi}_{Fi})}$ is the expected score for a person j and an item i, given $\tilde{\theta_j}$ (predicted score) and $\hat{\xi}_{Fi}$ (estimated item parameters for the focal group) from an item response model; $ES_{(ji|\tilde{\theta_j},\hat{\xi}_{Ri})}$ is the expected score for a person j and an item i, given $\tilde{\theta_j}$ (predicted score) and $\hat{\xi}_{Ri}$ (estimated item parameters for the reference group) from an item response model.

As shown in these equations, the differences between the two expected scores in the focal group are averaged across persons first and then summed across items in the $SIDS$ and the

$UIDS$. The main difference between the $SIDS$ and the $UIDS$ is that the $SIDS$ allows for full cancellation of DIF across persons and items, whereas the $UIDS$ allows no cancellation across persons or items (as absolute value of differences in the expected scores). Thus, large differences in the $|SIDS|$ and the $UIDS$ show that much cancellation takes place among items and/or persons implying non-uniform DIF, whereas similar $SIDS$ and $UIDS$ indicate that DIF is unidirectional (i.e., DIF favors one group uniformly).

Interpretation of the $STDS$ and the $UTDS$ is the difference in total scores, on average, across all persons in a focal group, due to DIF. For example, the $STDS$ of -1.0 indicates that, on average, the focal group would be expected to score 1.0 lower on the total score scale than the reference group with equal standing on the latent variable (when cancellation of DIF is allowed across persons and items).

Table 6 presents two scale-level DIF effect size measures, the $STDS$ and the $UTDS$, the values are on the total score scale (the two measures can range from 0 to 20) using one simulated data set for each simulation condition. In our simulation design, the $STDS$ and the $UTDS$ for uniform DIF type is higher than that for non-uniform DIF type for all conditions except for 10% DIF and low magnitude. As shown in Table 6, DIF was cancelled across both items and persons in all conditions of the non-uniform DIF type except 10% DIF and low magnitude. In both non-uniform and uniform DIF types, $UTSD$ increased as the number of DIF and the DIF magnitudes increased with one exception. The one exception is that the $UTDS$ value for 30% DIF items and high magnitude (1.003) is higher than that for 50% DIF items and low magnitude (0.810) in the uniform DIF type. For all DIF conditions except for 10% DIF and low magnitude in the non-uniform type, the $STDS$ has negative values in non-uniform and uniform DIF types, which indicates that on average the focal group is lower on the total score scale than the reference group. In the non-uniform DIF, there was larger cancellation as the number of DIF items and DIF magnitude increase.

Table 6: Scale-Level Effect Sizes for the Simulation Conditions

	Conditions			Effect Sizes		Cancellation
Condition Num.	Number	Type	Magnitude	$STDS$	$UTDS$	$\lvert STDS\rvert$-$UTDS$
1	10	Uniform	Low	-0.188	0.188	0
2	10	Uniform	High	-0.363	0.363	0
3	10	Non-uniform	Low	0.233	0.233	0
4	10	Non-uniform	High	-0.162	0.324	-0.162
5	30	Uniform	Low	-0.526	0.526	0
6	30	Uniform	High	-1.003	1.003	0
7	30	Non-uniform	Low	-0.306	0.429	-0.123
8	30	Non-uniform	High	-0.266	0.809	-0.543
9	50	Uniform	Low	-0.810	0.810	0
10	50	Uniform	High	-1.555	1.555	0
11	50	Non-uniform	Low	-0.457	0.792	-0.335
12	50	Non-uniform	High	-0.452	1.538	-1.086

Appendix H. Simulation Study Result Hypotheses

We expected the following simulation results when the same generated datasets were used to fit for the deleting, ignoring, multigroup, and modeling DIF approaches. First, it is expected that the patterns in the results for balanced and unbalanced designs are similar regarding the number of DIF items, the DIF magnitude, and the type of DIF, despite the different bias and RMSE magnitudes. In the multigroup and modeling DIF approaches, the magnitudes of the RMSE (as overall accuracy) for the item parameter estimates and the person scores are expected to be smaller in the unbalanced design than those in the balanced design because there are more persons in the reference group than in the focal group in the unbalanced design. However, in the multigroup (with the two-step) and modeling DIF approaches, the RMSEs for the impact and variance estimates (in the focal group) in the unbalanced design are expected to be larger than those in the balanced design because there is a smaller number of persons in the focal group than in the reference group in the unbalanced design. In the deleting and ignoring DIF approaches, the results are not expected to be affected by the sample size design because they are one-group analyses.

Second, we expected little bias for the multigroup DIF approach because the population data-generating model is the special case of the multigroup model. When the modeling DIF approach performs well, the accuracy for the modeling DIF approach is expected to be similar to that for the multigroup DIF approach and to be smaller than for the ignoring DIF approach.

Third, the bias is expected to be increasing in the ignoring DIF approach when there are a larger number of DIF items, higher DIF magnitudes, and uniform DIF. However, the accuracy is expected to be less affected by these simulation conditions in the multigroup and modeling approaches. It is because the effects of the DIF items increase with the larger number of DIF items and the higher DIF magnitudes (see the scale-level effect sizes in Table 6 in Appendix G). Overall, the uniform DIF type had a higher effect size than the non-uniform DIF type in our simulation design.

Fourth, it is expected that the RMSE for the modeling approach can be larger than that of the multigroup approach. Variance of the estimator tends to be larger when model complexity increases. Compared to the multigroup approach, the modeling approach has a larger number of parameters to be estimated. This pattern is more evident for the larger number of DIF items because the number of item discriminations to be estimated increases with the increasing number of DIF items.

Fifth, we expected that the standard errors of item parameter estimates from the multigroup DIF approach (with one-step approach) are lower than those from the modeling DIF approach because the standard errors from the multigroup DIF approach are calibrated using the reference group. Thus, the average ratio of SE_{MG} to SE_{M} is expected to be higher than 1.0.

Sixth, IRT reliability and accuracy of scores for the deleting approach are expected to be lower than those of the other three approaches because only non-DIF items are calibrated in the deleting DIF treatment (i.e., a smaller number of items is used to obtain the scale score). The IRT reliability for the multigroup approach with the two-step is expected to be similar to that of the modeling approach.

Appendix I. Simulation Study Results for an Unbalanced Design

Table 7: Average Bias, RMSE, and Ratio across Items for Item Parameters: Item Discrimination Parameter

		Uniform					Non-Uniform			
No. of DIF Items	Magnitude	Deleting	Ignoring	Multigroup	Modeling		Deleting	Ignoring	Multigroup	Modeling
						A. Bias				
10%	Low	0.033	0.037	0.012	0.012		0.036	0.019	0.014	0.017
	High	0.035	0.041	0.013	0.013		0.034	0.025	0.013	0.014
30%	Low	0.036	0.047	0.013	0.012		0.037	0.023	0.013	0.015
	High	0.037	0.056	0.013	0.011		0.037	0.011	0.013	0.015
50%	Low	0.040	0.058	0.013	0.012		0.044	0.016	0.014	0.013
	High	0.040	0.080	0.012	0.009		0.042	0.000	0.013	0.010
						B. RMSE				
10%	Low	0.089	0.090	0.081	0.084		0.092	0.097	0.086	0.093
	High	0.091	0.094	0.083	0.088		0.090	0.090	0.083	0.084
30%	Low	0.098	0.097	0.082	0.089		0.099	0.087	0.086	0.085
	High	0.099	0.107	0.084	0.099		0.099	0.088	0.087	0.086
50%	Low	0.110	0.103	0.082	0.093		0.111	0.085	0.090	0.084
	High	0.110	0.133	0.084	0.112		0.110	0.086	0.090	0.088
						C. Ratio				
10%	Low	-	-	-	1.001		-	-	-	1.004
	High	-	-	-	1.003		-	-	-	1.007
30%	Low	-	-	-	1.003		-	-	-	1.033
	High	-	-	-	1.011		-	-	-	1.026
50%	Low	-	-	-	1.003		-	-	-	1.056
	High	-	-	-	1.015		-	-	-	1.060
Aggregated Bias										
No. of DIF items	10%	0.035	0.031	0.013	0.014					
	30%	0.037	0.034	0.013	0.013					
	50%	0.042	0.039	0.013	0.011					
Magnitude	Low	0.038	0.033	0.013	0.013					
	High	0.037	0.036	0.013	0.012					
Type	Uniform	0.037	0.053	0.013	0.012					
	Non-Uniform	0.038	0.016	0.013	0.014					
Aggregated RMSE										
No. of DIF items	10%	0.090	0.093	0.083	0.087					
	30%	0.099	0.095	0.085	0.090					
	50%	0.110	0.102	0.086	0.094					
Magnitude	Low	0.100	0.093	0.085	0.088					
	High	0.100	0.099	0.085	0.093					
Type	Uniform	0.099	0.104	0.083	0.094					
	Non-Uniform	0.100	0.089	0.087	0.087					
Aggregated Ratio										
No. of DIF items	10%	-	-	-	1.004					
	30%	-	-	-	1.018					
	50%	-	-	-	1.034					
Magnitude	Low	-	-	-	1.017					
	High	-	-	-	1.020					
Type	Uniform	-	-	-	1.006					
	Non-Uniform	-	-	-	1.031					

Table 8: Average Bias, RMSE, and Ratio across Items for Item Parameters: Item Location Parameter

		Uniform					Non-Uniform			
No. of DIF Items	Magnitude	Deleting	Ignoring	Multigroup	Modeling		Deleting	Ignoring	Multigroup	Modeling
						A. Bias				
10%	Low	0.161	0.176	0.024	0.022		0.160	0.144	0.022	0.025
	High	0.160	0.186	0.024	0.022		0.161	0.183	0.025	0.022
30%	Low	0.171	0.201	0.023	0.026		0.171	0.192	0.023	0.025
	High	0.170	0.234	0.022	0.027		0.171	0.220	0.023	0.026
50%	Low	0.189	0.226	0.024	0.032		0.189	0.211	0.022	0.030
	High	0.188	0.283	0.022	0.035		0.191	0.259	0.023	0.036
						B. RMSE				
10%	Low	0.174	0.187	0.067	0.069		0.173	0.165	0.067	0.071
	High	0.173	0.198	0.067	0.075		0.174	0.195	0.068	0.072
30%	Low	0.184	0.212	0.068	0.079		0.184	0.203	0.068	0.077
	High	0.184	0.244	0.068	0.096		0.185	0.231	0.069	0.097
50%	Low	0.200	0.234	0.069	0.087		0.201	0.220	0.070	0.084
	High	0.200	0.290	0.069	0.105		0.202	0.267	0.070	0.111
						C. Ratio				
10%	Low	-	-	-	1.004		-	-	-	1.005
	High	-	-	-	1.000		-	-	-	1.002
30%	Low	-	-	-	1.017		-	-	-	1.024
	High	-	-	-	1.005		-	-	-	1.012
50%	Low	-	-	-	1.026		-	-	-	1.041
	High	-	-	-	1.013		-	-	-	1.025
Aggregated Bias										
No. of DIF items	10%	0.161	0.172	0.024	0.023					
	30%	0.171	0.212	0.023	0.026					
	50%	0.189	0.245	0.023	0.033					
Magnitude	Low	0.174	0.192	0.023	0.027					
	High	0.174	0.227	0.023	0.028					
Type	Uniform	0.173	0.217	0.023	0.027					
	Non-Uniform	0.174	0.202	0.023	0.027					
Aggregated RMSE										
No. of DIF items	10%	0.173	0.186	0.067	0.072					
	30%	0.184	0.222	0.068	0.087					
	50%	0.201	0.253	0.070	0.097					
Magnitude	Low	0.186	0.204	0.068	0.078					
	High	0.186	0.237	0.069	0.093					
Type	Uniform	0.186	0.227	0.068	0.085					
	Non-Uniform	0.186	0.213	0.069	0.085					
Aggregated Ratio										
No. of DIF items	10%	-	-	-	1.002					
	30%	-	-	-	1.015					
	50%	-	-	-	1.026					
Magnitude	Low	-	-	-	1.019					
	High	-	-	-	1.009					
Type	Uniform	-	-	-	1.011					
	Non-Uniform	-	-	-	1.018					

Table 9: Average Bias and RMSE across Persons for Person Scores and IRT Reliability

		Uniform					Non-Uniform			
No. of DIF Items	Magnitude	Deleting	Ignoring	Multigroup	Modeling		Deleting	Ignoring	Multigroup	Modeling
						A. Bias				
10%	Low	0.137	0.137	0.008	0.007		0.137	0.137	0.034	0.027
	High	0.137	0.137	-0.000	-0.001		0.137	0.137	0.005	0.006
30%	Low	0.137	0.137	-0.010	-0.010		0.137	0.137	-0.002	-0.001
	High	0.137	0.137	-0.034	-0.030		0.137	0.137	-0.021	-0.015
50%	Low	0.137	0.137	-0.026	-0.025		0.137	0.137	-0.012	-0.011
	High	0.137	0.137	-0.068	-0.052		0.137	0.137	-0.040	-0.031
						B. RMSE				
10%	Low	0.449	0.431	0.410	0.411		0.450	0.435	0.412	0.412
	High	0.449	0.430	0.411	0.412		0.449	0.433	0.413	0.413
30%	Low	0.478	0.431	0.413	0.414		0.478	0.435	0.416	0.416
	High	0.477	0.434	0.422	0.422		0.478	0.442	0.426	0.423
50%	Low	0.509	0.432	0.419	0.420		0.509	0.438	0.420	0.421
	High	0.509	0.441	0.441	0.435		0.509	0.453	0.440	0.435
						C. Reliability				
10%	Low	0.774	0.799	0.853	0.797		0.775	0.795	0.850	0.793
	High	0.774	0.800	0.853	0.797		0.774	0.796	0.851	0.794
30%	Low	0.731	0.801	0.854	0.799		0.732	0.796	0.851	0.795
	High	0.732	0.803	0.856	0.799		0.732	0.793	0.851	0.792
50%	Low	0.674	0.803	0.856	0.802		0.675	0.795	0.851	0.794
	High	0.674	0.808	0.860	0.798		0.674	0.791	0.851	0.791
Aggregated Bias										
No. of DIF items	10%	0.137	0.137	0.012	0.010					
	30%	0.137	0.137	-0.017	-0.014					
	50%	0.137	0.137	-0.036	-0.030					
Magnitude	Low	0.137	0.137	-0.001	-0.002					
	High	0.137	0.137	-0.026	-0.021					
Type	Uniform	0.137	0.137	-0.022	-0.019					
	Non-Uniform	0.137	0.137	-0.006	-0.004					
Aggregated RMSE										
No. of DIF items	10%	0.449	0.432	0.411	0.412					
	30%	0.478	0.435	0.419	0.419					
	50%	0.509	0.441	0.430	0.428					
Magnitude	Low	0.479	0.434	0.415	0.416					
	High	0.479	0.439	0.425	0.423					
Type	Uniform	0.479	0.433	0.419	0.419					
	Non-Uniform	0.479	0.439	0.421	0.420					
Aggregated Reliability										
No. of DIF items	10%	0.774	0.798	0.852	0.795					
	30%	0.732	0.798	0.853	0.796					
	50%	0.674	0.799	0.855	0.796					
Magnitude	Low	0.727	0.798	0.853	0.797					
	High	0.727	0.799	0.854	0.795					
Type	Uniform	0.727	0.802	0.855	0.799					
	Non-Uniform	0.727	0.794	0.851	0.793					

Note.
Results for multigroup approach were based on the two-step approach to compare results between the reference and focal groups.;
Bias and RMSE for the multigroup approach were based on MLE for the comparison with deleting and ignoring approaches.;
Reliability for the multigroup (with the two-step) and modeling approaches was calculated based on Bayes estimation.

Table 10: Bias and RMSE for Impact (top) and Variance (bottom)

		Uniform			Non-Uniform	
No. of DIF Items	Magnitude	Multigroup	Modeling		Multigroup	Modeling
				A. Bias		
10%	Low	-0.018	-0.021		0.086	0.060
	High	-0.050	-0.054		-0.033	-0.028
30%	Low	-0.088	-0.090		-0.056	-0.054
	High	-0.186	-0.167		-0.132	-0.111
50%	Low	-0.152	-0.151		-0.095	-0.094
	High	-0.320	-0.259		-0.208	-0.175
				B. RMSE		
10%	Low	0.026	0.033		0.088	0.065
	High	0.054	0.059		0.039	0.038
30%	Low	0.091	0.094		0.060	0.060
	High	0.187	0.170		0.133	0.114
50%	Low	0.153	0.154		0.097	0.097
	High	0.321	0.262		0.209	0.178
Aggregated Bias						
No. of DIF items	10%	-0.004	-0.011			
	30%	-0.115	-0.106			
	50%	-0.194	-0.170			
Magnitude	Low	-0.054	-0.058			
	High	-0.155	-0.132			
Type	Uniform	-0.136	-0.124			
	Non-Uniform	-0.073	-0.067			
Aggregated RMSE						
No. of DIF items	10%	0.052	0.049			
	30%	0.118	0.110			
	50%	0.195	0.173			
Magnitude	Low	0.086	0.084			
	High	0.157	0.137			
Type	Uniform	0.139	0.129			
	Non-Uniform	0.104	0.092			

Note. Impact for the multigroup approach was estimated with the two-step approach.

		Uniform			Non-Uniform	
No. of DIF Items	Magnitude	Multigroup	Modeling		Multigroup	Modeling
				A. Bias		
10%	Low	-0.033	-0.013		-0.101	-0.069
	High	-0.046	-0.017		-0.136	-0.065
30%	Low	-0.031	-0.007		-0.180	-0.147
	High	-0.064	-0.021		-0.322	-0.268
50%	Low	0.004	0.010		-0.238	-0.213
	High	0.023	0.044		-0.435	-0.369
				B. RMSE		
10%	Low	0.063	0.064		0.113	0.092
	High	0.070	0.066		0.144	0.090
30%	Low	0.064	0.066		0.186	0.159
	High	0.082	0.068		0.324	0.273
50%	Low	0.057	0.069		0.242	0.220
	High	0.066	0.088		0.436	0.374
Aggregated Bias						
No. of DIF items	10%	-0.079	-0.041			
	30%	-0.149	-0.111			
	50%	-0.161	-0.132			
Magnitude	Low	-0.097	-0.073			
	High	-0.163	-0.116			
Type	Uniform	-0.025	-0.001			
	Non-Uniform	-0.235	-0.189			
Aggregated RMSE						
No. of DIF items	10%	0.097	0.078			
	30%	0.164	0.141			
	50%	0.200	0.188			
Magnitude	Low	0.121	0.112			
	High	0.187	0.160			
Type	Uniform	0.067	0.070			
	Non-Uniform	0.241	0.201			

Note. Impact for the multigroup approach was estimated with the two-step approach.

References

Bingenheimer, J. B., Raudenbush, S. W., Leventhal, T., & Brooks-Gunn, J. (2005). Measurement equivalence and differential item functioning in family psychology. *Journal of Family Psychology, 19,* 441-455.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the psychopathy checklist-revised. *Psychological Assessment, 16,* 155-168.

Chiesi, F., Ciancaleoni, M., Galli, S., & Primi, C. (2012). Using the advanced progressive matrices (set I) to assess fluid ability in a short time frame: An item response theory-based analysis. *Psychological Assessment, 24,* 892-900.

Cho, S.-J., Wilmer, J., Herzmann, G., McGugin, R., Fiset, D., Van Gulick, A. E., Ryan, K., & Gauthier, I. (2015). Item response theory analyses of the Cambridge face memory test (CFMT). *Psychological Assessment, 27,* 552-566.

Church, A. T., Alvarez, J. M., Mai, N. T. Q., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the revised NEO personality inventory. *Journal of Personality and Social Psychology, 101,* 1068-1089.

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology, 85,* 451-461.

Cooke, D. J. Kosson, D. S., & Michie, C. (2001). Psychopathy and ethnicity: Structural, item, and test generalizability of the Psychopath Checklist-Revised (PCL-R) in Caucasian and African American participants. *Psychological Assessment, 13,* 531-542.

DeWalt, D., Thissen, D., Stucky, B. D., Langer, M. M., Morgan DeWitt, E., Irwin, D. E., Lai, J.-S., Yeatts, K. B., Gross, H. E., Taylor, O., & Varni, J. W. (2013). PROMIS pediatric peer relationships scale: Development of a peer relationships item bank as part of social health measurement. *Health Psychology, 32,* 1093-1103.

Donovan, M. A., Drasgow, F., & Probst, T. M. (2000). Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *Journal of Applied Psychology, 85,* 305-313.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44,* 576-585.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74,* 912-921.

Ellis, B. B., & Kimmel, H. D., (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77,* 177-184.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology, 86,* 215-227.

Fledderus, M., Oude Voshaar, M. A. H., ten Klooster, P. M., & Bohlmeijer, E. T. (2012). Further evaluation of the psychometric properties of the acceptance and action questionnaire-II. *Psychological Assessment, 24,* 925-936.

Freeman, A. J., Youngstrom, E. A., Frazier, T. W., Youngstrom, J. K., Demeter, C., & Findling, R. L. (2012). Portability of a screener for pediatric bipolar disorder to a diverse setting. *Psychological Assessment, 24,* 341-351.

Kim, E.-S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18,* 212-228.

McCarthy, D. M., Pedersen, S. L., & D'Amico, E. J. (2009). Analysis of item response and differential item functioning of alcohol expectancies in middle school youths. *Psychological Assessment, 21,* 444-449.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95,* 728-743.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97,* 1016-1031.

Meulders, M. & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach.* (pp. 213-240). New York, NY: Springer.

Mungas, D., Reed, B. R., Crane, P. K., Haan, M. N., & González, H. (2004). Spanish and English neuropsychological assessment scales (SENAS): Further development and psychometric characteristics. *Psychological Assessment, 16,* 347-359.

Neal, D. J., Corbin, W. R., & Fromme, K. (2006). Measurement of alcohol-related consequences among high school and college students: Application of item response models to the Rutgers alcohol problem index. *Psychological Assessment, 4,* 402-414.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96,* 966-980.

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish Translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment, 14,* 50-59.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517-529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552-566.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the multidimensional personality questionnaire stress reaction scale. *Journal of Personality and Social Psychology, 75,* 1350-1362.

Stark, S., Chernyshenko, O. S., Chan, K.-Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86,* 943-953.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89,* 497-508.

Whitney, D. J., & Schmitt, N. (1997). Relationship between culture and responses to biodata employment items. *Journal of Applied Psychology, 82,* 113-129.

Wiesner, M., Chen, V., Windle, M., Elliott, M. N., Grunbaum, A., Kanouse, D. E., & Schuster, M. A. (2010). Factor structure and psychometric properties of the brief symptom inventory- 18 in women: A MACS approach to testing for invariance across racial/ethnic groups. *Psychological Assessment, 22,* 912-922.

Wu, J., King, K. M., Witkiewitz, K., Racz, S. J., & McMahon, R. J. (2012). Item analysis and differential item functioning of a brief conduct problem screen. *Psychological Assessment, 24,* 444-454.