Psychological Test and Assessment Modeling, Volume 57, 2015 (3), 279-300

Modeling DIF for simulations: Continuous or categorical secondary trait?

Christine E. DeMars[1]

Abstract

For DIF studies, item responses are often simulated using a unidimensional item response theory (IRT) model, with the item difficulty varying by group for the DIF item(s). This implies that the item is easier for all members of one group. However, many researchers may want to conceptualize DIF as a continuous factor. In this conceptualization, one group has higher average scores on the factor causing the DIF, but there is variance within groups. Multidimensional IRT models allow item responses to be generated to correspond to this perspective. Data were simulated under both unidimensional and multidimensional models and effect sizes were estimated from the resulting item responses using the Mantel-Haenszel DIF procedure. The bias and empirical standard errors of the effect sizes were virtually identical. Thus, practitioners using observed-score methods of DIF detection can trust results from DIF simulation studies regardless of the underlying model used to create the data.

Keywords: DIF, IRT, Mantel-Haenszel

[1] *Correspondence concerning this article should be addressed to:* Christine E. DeMars, PhD, James Madison University, Center for Assessment & Research, MSC 6806, 298 Port Republic Road, Harrisonburg, Virginia 22807, USA; email: demarsce@jmu.edu

In studies of differential item functioning (DIF), researchers study whether the log-odds (or sometimes probability) of correct response to an item is the same for two groups after controlling for differences in ability. One group is termed the reference group and one is termed the focal group. If there are conceptual or policy reasons for concern with one group's performance, that group is labelled the focal group. Otherwise, the distinction is arbitrary. Simulations are frequently used when developing new models or estimators for DIF analyses. With real data, researchers can only determine whether two estimators are different, but with simulations they can determine which is more accurate because the true values are known.

For simulation studies, DIF is often modelled in an IRT framework as a difference in item parameters. The secondary trait underlying the DIF is thus categorical, applying to all group members. The shift in the item response function is the same for every member within the focal group or within the reference group. For example, if the discrimination parameter is the same for both groups but the item difficulty is higher for the focal group, the item will be harder for all focal group members than for matched reference group members. However, it might seem more reasonable to conceptualize the item parameter difference as representing the *average* DIF, not a constant that applies to each group member. The response probabilities of some focal group members may be more like the response probabilities of reference group members, and the other way around. Some have proposed applying IRT mixture models to account for this (Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002). In mixture models, the reference and focal group may be disproportionally distributed across latent classes, but some members of each group may have a high posterior probability of membership in the class disadvantaged by the DIF item. However the use of latent class models implies that the DIF changes the response function in the same way for all members of a class. DIF might be more realistically conceptualized as a continuous secondary trait that is distributed differently in the reference and focal groups (Ackerman, 1992; Camilli, 1992; Roussos & Stout, 1996). Another equivalent way of phrasing this is that DIF could be considered a random effect.

This article examines whether simulating DIF as a continuous secondary trait (random effect) using a multidimensional model is comparable to simulating DIF as lack of invariance in the unidimensional item parameters in terms of the effect size estimated by observed-scored indices. In the DIF literature, the most common indices, such as the Mantel-Haenszel (Dorans, 1989; Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), and SIBTEST (Shealy & Stout, 1993), do not use IRT. Instead, they condition on observed score or on a true score based on classical test theory (CTT). However, simulated data for studying observed-score indices is typically generated with IRT models, and the choice of a continuous or categorical secondary trait for the data generation may have implications for the findings for indices conditioned on observed score.

Most simulation studies have generated the DIF as a categorical trait, but some scholars have argued that theoretically a continuous secondary trait causes DIF. Ackerman (1992) provided a detailed explanation of how group differences in the distribution of a secondary trait, termed a nuisance trait, would lead to DIF. Roussos and Stout (1996) and Ca-

milli (1992) each described further mathematical details. Several additional researchers have advocated this conceptualization (Bolt & Stout, 1996; Douglas, Roussos, & Stout, 1996). Additionally, Ackerman (1992), Jiang and Stout (1998), and Shealy and Stout (1993) simulated DIF using a continuous secondary trait and studied power for the MH and Sibtest procedures. Li and Stout (1996) also simulated DIF this way and included logistic regression.

The research question explored in this study was: Does simulating DIF with a MIRT model, compared to a unidimensional model with group-specific item difficulties, yield Mantel-Haenszel DIF effect size estimates that differ in bias or standard error? There was no reason to think that the bias would depend on the method of simulating DIF, but it seemed plausible that the estimates might be more stable across replications when the DIF was simulated with a unidimensional model because the difference in item difficulties had the same effect on all group or class members. If the MIRT conceptualization of DIF better represents real-life cognitive processes, the stability of the unidimensional method of simulation would be inaccurate. In study 1, responses to each DIF item were a function of the primary trait and a different item-specific secondary trait. In study 2, there were multiple clusters of DIF items. Responses to each of the items within a cluster were a function of the primary trait and a cluster-specific secondary trait.

Study 1

Study 1 – Method

Data simulation. Data were generated to follow either a 3PL or 1PL model. The unidimensional model was the 3PL in the normal metric or the 1PL model in the logistic metric. The multidimensional 3PL model was:

$$P(x=1) = c + (1-c)\frac{e^{1.7(a_p\theta_p + a_s\theta_s + d)}}{1+e^{1.7(a_p\theta_p + a_s\theta_s + d)}}, \quad (1)$$

where x is the item response, 0 if the response is incorrect and 1 if it is correct, c is the lower asymptote, a_p is the discrimination parameters parameter for the primary trait θ_p, a_s is the discrimination parameters parameter for the secondary DIF trait θ_s , and d is an easiness parameter[2]. In the undimensional model, item difficulty ($b = -d/a$) usually replaces d. To make the d-parameter more comparable to the unidimensional b parameter, the MID index (Reckase & McKinley, 1991) can be calculated:

$$MID = \frac{-d}{\sqrt{a_p^2 + a_s^2}}. \quad (2)$$

[2] Subscript i could be added to x, a_p, a_s, and d to indicate item i. Subscript j could be added to x, θ_p, and θ_s to indicate person j.

The multidimensional 1PL model was:

$$P(x=1)=\frac{e^{(\theta_p+\theta_s+d)}}{1+e^{(\theta_p+\theta_s+d)}}, \quad (3)$$

with all terms as defined for Equation 1. Whereas the a_s controls the amount of DIF for the 2PL or 3PL models, the variance of θ_s controls the amount of DIF in the multidimensional 1PL model. As $\sigma_{\theta s}$ increases, the logit difference between the groups increases. Alternatively, the θ variances could be fixed to one and an *a*-parameter could be added for each trait. The *a*-parameter for the primary θ would be constant across items.

In each of 5000 replications, responses were simulated from 2000 reference group members and 2000 focal group members. The term *impact* is often used in the DIF literature to indicate true group mean differences in the primary ability, in contrast to item-specific group differences[3]. Impact has been shown to bias Mantel-Haenzel estimates unless the data follow a 1PL model (Li, Brooks, & Johanson, 2012; Wainer & Skorupski, 2005; Zwick, 1990). It was included here to see if the effect of impact changed when the DIF was due to a continuous secondary trait. There were two levels of impact in this study: The mean of the primary trait was 0 for both groups in the no-impact condition, but 0.5 for the reference group and -0.5 for the focal group in the impact condition. The within-group standard deviation was one.

The simulated test had 60 items, 12 with DIF and 48 DIF-free. The DIF was balanced, with six items favoring the focal group and six favoring the reference group.

For the 3PL data, the DIF-free items had four levels of *a*-parameters (0.9, 1.1, 1.3, 1.5) crossed with 12 levels of *b*-parameter (ranging from -2.14 to 2.14, spaced to reflect a normal density). The *c*-parameters were 0.2 for both DIF and DIF-free items. For the 1PL data, *b*-parameters were selected to minimize the difference between the item response function (IRF) for the 3PL data and the IRF for the 1PL data. In the minimization, the points on the IRF were weighted by a standard normal density so that the predicted probabilities would be most similar in the regions where there were more examinees. The resulting *b*-parameters ranged from -3.33 to 1.73. The parameters are available in Table A1 in Appendix A.

DIF magnitude was indexed by the Δ-difference, an effect size measure popularized by the Educational Testing Service. The Δ-difference = -2.35 times the mean difference in log-odds (Dorans, 1989; Holland & Thayer, 1985). Typically, items are classified as A (little or no DIF) if the Δ-difference is not statistically significantly different from 0 and/or < 1 in absolute value, B if the Δ-difference is significantly different from 0 and > 1 in absolute value, and C if the Δ-difference is significantly different from 1 and > 1.5 in absolute value (Zieky, 1993).

[3] The term *impact* is used in the standard-setting literature to simply mean group differences in total scores, because if the scores are used to make decisions about students, those decisions will *impact* the groups of students differently if they have different mean scores.

First, continuous DIF was simulated using Equation 1 or 3. A different secondary trait influenced each DIF item, consistent with the definition of DIF as a disturbance which influences isolated items (De Boeck, 2008; Thissen, Steinberg, & Gerrard, 1986), not an unintended substantive trait measured by a cluster of items[4]. Within groups, the secondary traits were uncorrelated with the primary trait or each other. Both the primary and secondary traits were normally distributed, with a within-group standard deviation of one for the 3PL data. The within-group standard deviation of the secondary trait was 0.7 or 0.5 for the 1PL data. The mean of the secondary trait was 0.5 for the reference group and -0.5 for the focal group for the first six DIF items, with the means reversed for the remaining six DIF items. For the 3PL data, three levels of item difficulty were crossed with two levels of a_p. Similar to the DIF-free items, the item difficulties for the 1PL model were chosen to minimize the difference between the 3PL and 1PL IRFs. The a_s parameter for the 3PL model or the within-group $\sigma_{\theta s}$ for the 1PL model was selected to yield Δ-differences[5] near the B/C border (for easy and middle difficulty items) or near the A/B border (for difficult items). In unidimensional DIF models, when there is a non-zero lower asymptote, it takes a larger *b*-difference to yield a given Δ-difference as item difficulty increases. Similarly, using the multidimensional model it takes a larger a_s parameter (or, equivalently, greater variance in θ_s) to produce a given Δ-difference as item difficulty increases, if the lower asymptote is not zero. The item parameters and Δ-differences for items 1-6 are shown in the left side of Tables 1 and 2. The parameters were the same for items 7-12 in the MIRT model, with only the means of the secondary traits reversed so that items 7-12 favored the focal group.

Table 1:
Parameters for 3PL DIF Items

	MIRT				unidimensional *b*-difference				
Item	a_1	a_2	MID	Δ-diff	*a*	b_{ref}	b_{foc}	*c*	Δ-diff
1	0.7	0.45	-1.5	-1.52	0.65	-2.03	-1.39	0.23	-1.51
2	1.5	0.50	-1.5	-1.58	1.34	-1.74	-1.40	0.21	-1.56
3	0.7	0.60	0.0	-1.56	0.58	-0.44	0.42	0.20	-1.55
4	1.5	0.70	0.0	-1.53	1.19	-0.24	0.23	0.20	-1.52
5	0.7	0.90	1.5	-1.12	0.50	1.76	3.05	0.19	-1.12
6	1.5	1.60	1.5	-1.04	0.77	1.63	2.70	0.19	-1.04

Note: In the MIRT model, all $c = 0.2$. For items 7-12, the Δ-differences were positive. Values for the MIRT Δ-difference and all values for the unidimensional model changed very slightly when there was a group mean difference in the primary θ. Values used in calculations had greater precision (0.0000001) then the values shown here.

[4] Study 2 examines clusters of items.

[5] True Δ-differences were calculated using the procedures of Roussos, Schnipke, and Pashley (1999), after marginalizing the response function over the secondary trait.

Table 2:
Parameters for 1PL DIF Items

	MIRT			unidimensional *b*-difference	
Item	$\sigma_{\theta 2}$	MID	Δ-diff	b_{ref}	b_{foc}
1	0.7	-1.87	-1.54	-2.20	-1.55
2	0.7	-2.18	-1.55	-2.51	-1.85
3	0.7	-0.40	-1.50	-0.72	-0.08
4	0.7	-0.36	-1.50	-0.68	-0.04
5	0.5	1.08	-1.12	0.84	1.32
6	0.5	1.31	-1.12	1.07	1.55

Note: For items 7-12, the Δ-differences were positive. Values for the MIRT Δ-difference changed very slightly when there was a group mean difference in the primary θ. Values used in calculations had greater precision (0.0000001) then the values shown here.

The categorical DIF was created from a unidimensional model with separate *b*-parameters for each group. For the 1PL univariate DIF, one half of the log-odds difference (corresponding to the MIRT Δ-difference) was subtracted from the *b*-parameter for one group and added to the *b*-parameter for the other group. The 3PL univariate *b*-parameters were based on the marginal response function. The marginal response functions were calculated from the MIRT parameters by integrating over the secondary θ distribution in each group. An example is shown in Figure 1. In the left panel, the item response surface is the same for both groups, but the marginal response function in the right panel differs by group. The unidimensional item parameters which best matched the marginal response functions, subject to the constraint of equal *a* and *c* parameters across groups, were found using the nlm function in the *stats* package in R (2014). For these calculations, each θ distribution was approximated using 50 quadrature points from – 4 to 4. Because the marginal response functions did not perfectly fit a parametric unidimensional model, the best approximation yielded slightly different Δ-differences for the 3PL data. The unidimensional item parameters and corresponding Δ-differences are also shown in Tables 1 and 2. After the best-fitting item parameters were established, data were generated using the unidimensional DIF model, with different *b*'s for each group. The primary θs and the 48 DIF-free item responses remained the same; only the DIF item responses were re-simulated.

In summary, the factors were process underlying the DIF (continuous/categorical) by response model form (3PL/1PL) by impact (mean difference = 0/1) for a 2 x 2 x 2 design. Recall that the research question focused on the effects of a continuous or categorical trait underlying the DIF. The other two factors have predictable effects in the context of a categorical DIF trait and were only of concern if they interacted with the continuous or categorical form of the DIF.

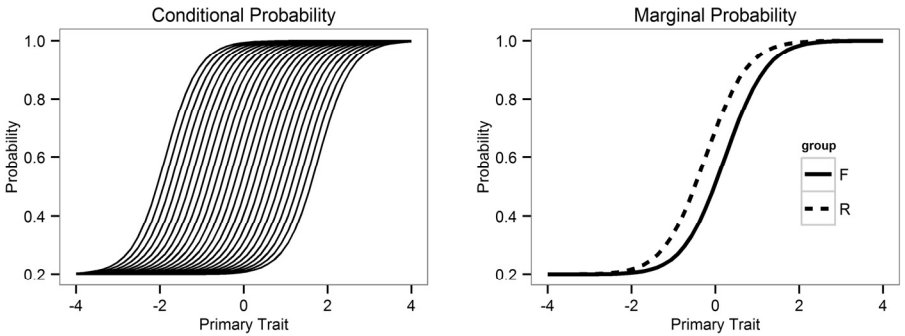

Figure 1:
Conditional and marginal probability. In the left panel, each contour line shows the response function at a different value of the secondary trait. In the right panel, the response function has been marginalized over a N(0.5,1) θ_s distribution for the reference group and a N(-0.5,1) θ_s distribution for the focal group.

Index of DIF magnitude. After generating data based on each model, the Δ-difference was estimated using the Mantel-Haenszel (MH) procedure (Dorans, 1989; Holland & Thayer, 1988). The logistic regression procedure estimates the same parameter, but treats the matching score as interval rather than nominal. In both procedures, the matching score can be either the total score on the test or the sum of the score on the studied item plus a subset of items believed to be DIF free. The DIF-free items may be chosen based on either empirical procedures or conceptual considerations. In the MH procedure, as described in Holland and Thayer (1988), examinees are divided into J subgroups, where J represents the number of possible summed scores on the matching test. The odds ratio α for the studied item is estimated as

$$\hat{\alpha} = \frac{\sum_{j=1}^{J} n_{Rj1} n_{Fj0} / T_j}{\sum_{j=1}^{J} n_{Rj0} n_{Fj1} / T_j}, \tag{3}$$

where n_{Rj1} is the number of reference group members at score *j* who answered the item correctly, n_{Rj0} is the number of reference group members at score *j* who answered the item incorrectly, n_{Fj1} is the number of focal group members at score *j* who answered the item correctly, n_{Fj0} is the number of focal group members at score *j* who answered the item incorrectly, and T_j is the total number of examinees who scored *j* on the matching test. The estimate of Δ is then $\hat{\Delta} = -2.35 \ln(\hat{\alpha})$. Negative values indicated DIF favoring the reference group. For this study, the score on the DIF-free items, plus DIF item *i* when item *i* was the item for which the index was calculated, was used as the matching variable, with each score (0-49) as a matching level. This represented a situation in which the

anchor items had been previously screened for DIF, or a purification process had accurately removed the other five DIF items.

Study 1 – Results

Accuracy of the Δ-difference was indexed using bias (mean difference, across replications, between estimate and true parameter) and standard error (standard deviation of the difference between estimate and mean estimate). The MSE, not reported in the figures and tables due to redundancy, would be the sum of the squared bias and squared standard error. For the DIF-free items, to serve as a baseline, Figure 2 shows the estimated bias and Figure 3 shows the estimated standard error. Because the DIF items were not used in the matching score when examining DIF-free items, it made no difference whether the DIF was continuous or categorical and this factor is not shown in the figures.

When there was no impact (no group mean difference in the primary trait), bias in the Δ-difference was very close to zero. The empirical standard errors were larger for the easier items, and conditional on difficulty the standard error increased with item discrimination for the 3PL items. The effect of discrimination and difficulty decreased as difficulty increased, with a small increase in standard error for the most difficult items.

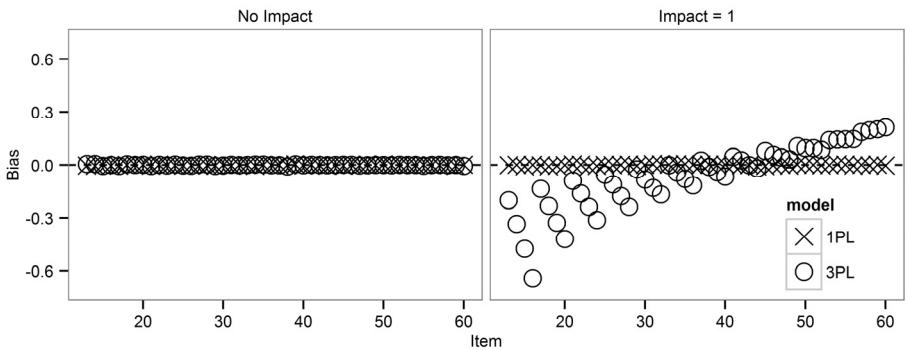

Figure 2:
Bias of Δ-difference for DIF-free items. The x-axis begins at item 13; information on items 1-12 appears in Table 3.Each set of four DIF-free items had the same difficulty in the 3PL model, with increasing discrimination. The difficulty (MID) increased for the next set of items.

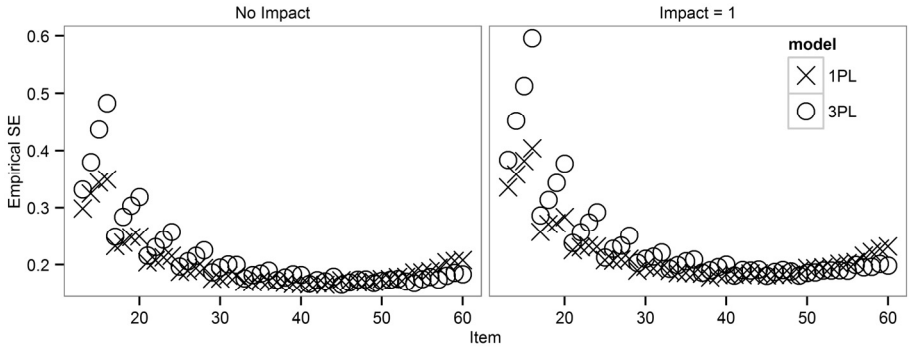

Figure 3:
Empirical standard error of Δ-difference for DIF-free items. The x-axis begins at item 13; information on items 1-12 appears in Table 4.

When there was impact, the bias depended on the item difficulty and discrimination, as well as whether the data were 3PL or 1PL. This phenomena was expected based on theory and past research and is discussed further in Appendix B.

The bias and standard errors in the Δ-difference are shown in Tables 3 and 4 for the DIF items. The bias was small and would have minimal effect considering the scale of the Δ-difference. As was found for the DIF-free items, when there was impact, the degree of

Table 3:
Bias of the Δ-difference for DIF Items

	3PL				1PL			
	No Impact		Impact = 1		No Impact		Impact = 1	
Item	continuous	categorical	continuous	categorical	continuous	categorical	continuous	categorical
1	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00
2	0.06	0.06	-0.28	-0.27	-0.01	-0.01	0.01	-0.01
3	-0.01	-0.01	0.07	0.07	-0.01	0.00	0.02	0.00
4	0.01	0.01	-0.02	-0.02	-0.01	0.00	0.02	0.00
5	0.00	0.00	0.18	0.18	-0.01	0.00	0.01	0.00
6	0.01	0.00	0.18	0.18	0.00	0.00	0.01	-0.01
7	0.01	0.00	-0.02	-0.01	0.01	0.00	0.02	0.01
8	-0.05	-0.05	-0.42	-0.41	0.01	0.00	0.02	0.00
9	0.02	0.01	0.09	0.09	0.00	0.00	0.03	0.00
10	-0.01	-0.01	-0.06	-0.06	0.01	0.00	0.03	0.00
11	0.01	0.00	0.19	0.19	0.00	0.00	0.01	0.00
12	-0.01	0.00	0.18	0.18	0.01	0.00	0.02	0.00

Table 4:
Empirical Standard Error of the Δ-difference for DIF Items

	3PL				1PL			
	No Impact		Impact = 1		No Impact		Impact = 1	
Item	continuous	categorical	continuous	categorical	continuous	categorical	continuous	categorical
1	0.24	0.24	0.27	0.27	0.23	0.22	0.27	0.24
2	0.32	0.32	0.38	0.38	0.26	0.24	0.30	0.27
3	0.16	0.17	0.18	0.18	0.17	0.17	0.18	0.18
4	0.18	0.18	0.19	0.19	0.17	0.17	0.18	0.18
5	0.16	0.17	0.18	0.18	0.18	0.18	0.20	0.20
6	0.17	0.17	0.19	0.19	0.19	0.19	0.21	0.21
7	0.24	0.24	0.26	0.26	0.24	0.22	0.26	0.25
8	0.32	0.32	0.36	0.36	0.27	0.24	0.29	0.26
9	0.17	0.16	0.19	0.19	0.17	0.17	0.19	0.19
10	0.18	0.18	0.20	0.20	0.17	0.17	0.19	0.19
11	0.17	0.17	0.19	0.19	0.19	0.18	0.20	0.20
12	0.17	0.17	0.20	0.20	0.19	0.19	0.21	0.21

bias in the 3PL estimates depended on the item parameters. Again, see Appendix B for a further discussion. The most important message from these tables is that bias and empirical standard error were nearly identical across the unidimensional and multidimensional conditions. The secondary dimension influenced examinees to varying extents within groups when DIF was modeled with MIRT. In contrast, the unidimensional model changed the *b*-parameter to the same degree for all group members. Importantly, this difference did not change the stability of the DIF estimates. The marginal response functions, which had a lower *a*-parameter after integrating across the secondary dimension, adequately captured the random variance within groups. The only exceptions were the easiest items in the 1PL data. These items had somewhat higher SEs when the DIF was due to a continuous secondary dimension, sometimes as much as 10% higher.

Study 2

Study 2 – Method

In some contexts, it may make sense to model multiple DIF items using the same secondary trait. This was explored using the same methodology as Study 1, except that one secondary trait influenced a set of six DIF items, and another trait influenced another set of six DIF items, and so on. This is closer to the conceptualization of DIF used in IRT mixture models, except that the secondary trait in IRT mixture models is categorical. The

secondary traits were uncorrelated with the primary trait and with each other. This is not meant to imply that DIF would not be related to an auxiliary trait highly correlated with the primary trait. For example, on a math test, DIF might be related to reading or to a subskill within math, which would be highly correlated with the primary math proficiency. But the part of reading or the part of the subskill that causes the DIF is the residual part that is uncorrelated with the primary trait. The secondary trait modeled here is only this residual, not the entirety of the conceptual trait. This is equivalent to a bifactor model.

The research question remained: Does simulating DIF with a MIRT model, compared to a unidimensional model with group-specific item difficulties, yield Mantel-Haenszel DIF effect size estimates that differ in bias or standard error?

The sample size was again 2000 examinees in the reference group and 2000 in the focal group, with each condition replicated 2000 times. The three factors from Study 1 were repeated: process underlying the DIF (continuous/categorical) by response model form (3PL/1PL) by impact (mean difference = 0/1).

Two additional factor were added for a 2 x 2 x 2 x 2 x 2 design: matching score and proportion of DIF items. There were two levels of matching score: DIF-free items plus studied item, or total number correct. There were three levels of number of DIF items: 12, 24, or 36 DIF items out of 60 total items (20%, 40%, 60%). For the 12 DIF item condition, the item parameters were identical to those in Study 1, but the DIF for items 1-6 was due to a single secondary trait and the DIF for items 7-12 was due to another secondary trait. For the 24 and 36 DIF item conditions, DIF was added to some of the formerly DIF-free items. Each set of 6 DIF items was influenced by a different secondary trait. To balance the DIF contamination in the total score as much as possible, the reference group scored higher on half the secondary traits and the focal group scored higher on the other half. For both the 24 and 36 DIF-items conditions, items 13, 24, 30, 38, 47, 55 were a function of a secondary trait favoring the reference group, and items 14, 23, 31, 37, 48, 54 were a function of a secondary trait favoring the focal group. Additionally for the 36 DIF-items condition, items 20, 25, 35, 44, 51, 57 favored the reference group and items 19, 26, 34, 43, 50, 58 favored the focal group. As in study 1, the difference in group means for the secondary trait was 1 standard deviation unit. The univariate parameters for the categorical DIF were then based on the response function after marginalizing over the secondary trait, as explained in Study 1. The parameters for the additional DIF items are shown in Appendix A.

Study 2 – Results

DIF-free items. When there was no impact (no group mean difference in the primary trait), bias in the Δ-difference was almost zero for all DIF-free items, regardless of the number of DIF items or the type of matching or whether the DIF was categorical or continuous or whether the data followed the 1PL or 3PL model.

For impact = 1, the bias in the Δ-difference for the DIF-free items is shown in Figure 4 for the 3PL items. When examinees were matched on observed score on the DIF-free items only, the bias in the easiest and most difficult 3PL items became more extreme as the number of DIF items increased because the number of items in the matching score, and thus the reliability of the matching score, decreased (see Appendix B for further explanation). This result was not observed when all items were used in the matching variable because the reliability of the scores did not change very much as DIF items replaced DIF-free items. However, pertinent to the research question, if all items were included in the matching variable, as the number of DIF items increased the bias in the easiest DIF-free items became slightly more negative when the DIF in the other items was modeled as continuous rather than categorical (lower right panel in Figure 4). Although consistent across all easy to moderate-difficulty items, the difference was very small (.01 or .02) and of no practical importance.

When the data followed a 1PL model, there was no bias in the Δ-difference for the DIF-free items in any condition when only the DIF-free items were used in the matching score or when the group mean impact was zero. However, when all items were used in the matching score and impact = 1, there was a small negative bias when the DIF was modeled as a continuous factor (random effect), which increased as the number of DIF items increased. Although the bias was consistent across items, it was very small: -.01 when there were 12 DIF items, -.03 when there were 24 DIF items, and -.05 when there

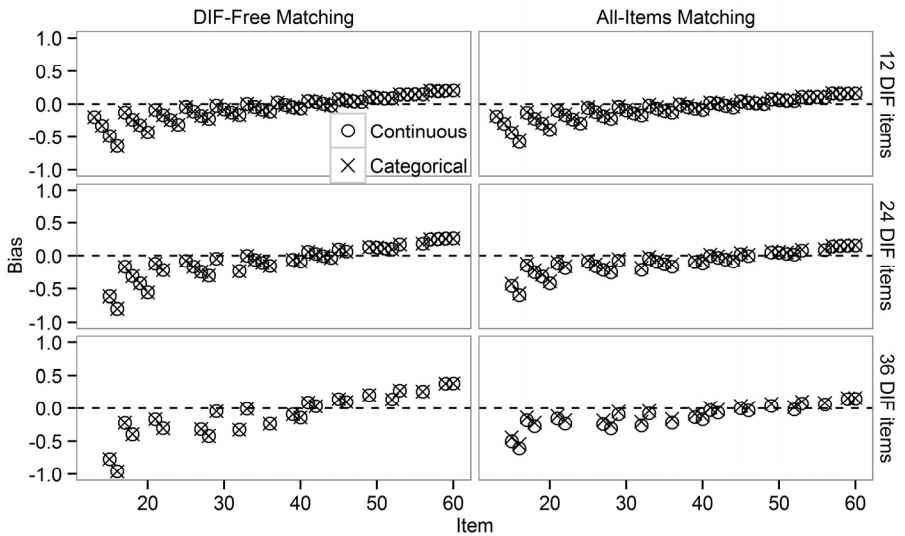

Figure 4:
Bias of Δ-difference for DIF-free 3PL items in Study 2 with impact = 1. The x-axis begins at item 13 because items 1-12 had DIF in all conditions. The y-axis covers a larger range than it did in Figure 2.

were 36 items. This occurred when the DIF was modeled as a continuous factor because the number correct score is not a sufficient statistic for θ under the multidimensional 1PL model. When there is group mean impact and some items in the matching score measure both factors, examinees matched on number correct score are not precisely matched on the underlying latent score. In this data, each group had a higher mean on half of the secondary θs, and thus on average neither group had higher total scores due to the DIF. Because the reference group had a higher mean on the primary θ, in a subgroup with the same number correct scores, the reference group had a slightly higher mean primary θ and the focal group had a slightly higher mean over the secondary θs. Thus, the DIF-free items which measured only the primary θ appeared easier for the reference group. However, the mismatch was far less than it was for the 3PL data. It would not be worth noting except that it is one of the few differences directly relevant to the research question of differences between categorical and continuous DIF.

The standard error of the Δ-difference for the DIF-free items was nearly identical to the values in Figure 3 in Study 1. It did not depend on the number of DIF items or which items were summed in the matching variable and was the same regardless of whether the other items had categorical or continuous DIF.

DIF items. The DIF items showed almost no bias in the Δ-difference when impact = 0. As in study 1, the easy, high-discrimination items showed a small amount of bias for the 3PL data (see Appendix B for an explanation). Additionally, when all items were included in the matching score, there was a very small negative bias (approximately -.02 to -.04) for items which favored the reference group and a similar very small positive bias for items which favored the focal group. This is illustrated in the left panel of Figure 5 for the 1PL data and was similar for the 3PL data. Bias this small would have no practical effect.

When impact = 1, the bias for the DIF items followed the same pattern seen in Study 1 as long as the DIF-free items (plus the studied item) were used as the matching score, with almost no difference between the Δ-difference estimates for the continuous and categorical DIF. When all items were used in the matching score for the 3PL data, the positive bias for difficult items became very slightly less positive for items that favored the reference group, and the negative bias for easy items became slightly less negative for items that favored the focal group. When all items were used in the matching score for the 1PL data, the bias was slightly more positive when the DIF was continuous rather than categorical (Figure 5). Items which favored the reference group had a slight negative bias (approximately -0.02 to -0.03) in the Δ-difference for the categorical DIF as they did when impact = 0. But these items had almost zero bias for the continuous DIF. Similarly, items which favored the focal group had a slight positive bias in the Δ-difference for the categorical DIF but a larger positive bias for the continuous DIF. Again, the difference between the estimates for the categorical and continuous DIF were too small to have much practical effect.

The explanation for this small difference between categorical and continuous DIF is that the continuous DIF was from a common cause for all items within a set. Individual examinees who were higher on the secondary trait had higher probabilities on all six items

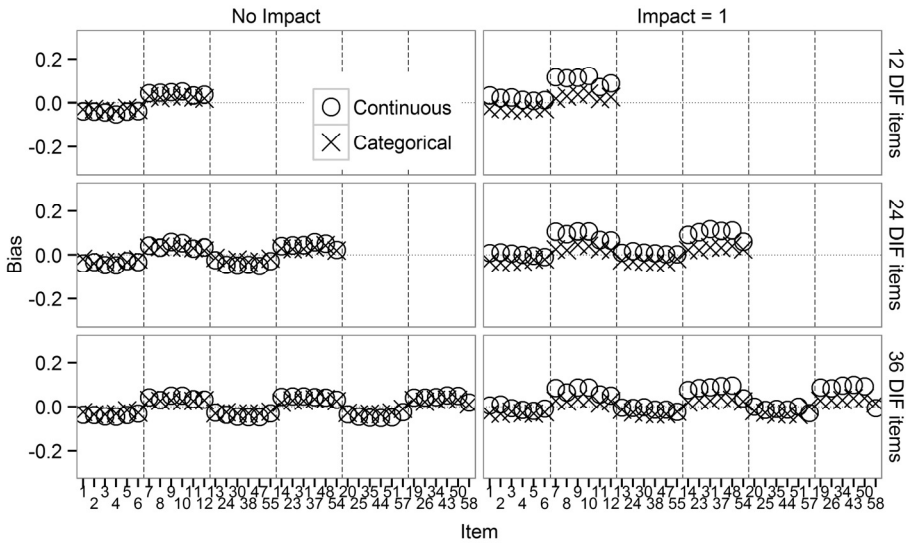

Figure 5:
Bias of Δ-difference for 1PL DIF items in Study 2 when all items were used in the matching score. Items are grouped into sets of six items with a common secondary trait in the MIRT model, with dashed vertical lines separating the item sets.The y-axis covers a smaller range than it did in Figures 2 or 4, so small differences are accentuated.

that tapped this trait, thus boosting their total observed scores beyond what would be expected from their primary θ. Because the secondary θ that influenced one group of six items was uncorrelated with the secondary θ that influenced the other groups of items, this bias in the total score did not balance out at the examinee level. In contrast, when DIF was modeled as a difference in *b*-parameter using a unidimensional model, all examinees had an increased probability on half the DIF items balanced by a corresponding decreased probability on the other half of the DIF items.

The standard error of the Δ-difference for the DIF items was again nearly identical to the values in Figure 3 in Study 1. It did not depend on the number of DIF items or which items were summed in the matching variable. For the 3PL data, it was the same regardless of whether the other items had categorical or continuous DIF. For the 1PL data, the standard error for the easiest items was somewhat larger when the DIF was due to a continuous trait, as in Study 1.

Limitations

The common limitations of simulation studies apply here. The data perfectly fit either the multidimensional or unidimensional 3PL or 1PL model. Only a limited number of factors were studied. Sample size and test length, for example, were not varied because their effects on DIF estimation using the MH procedure are well known and were not expected to vary depending on the model underlying the DIF. Unbalanced DIF, where more than half of the DIF items favor one group, was also not studied. With unbalanced DIF, it is important to use a purification procedure to yield a set of DIF-free items for the matching test. Otherwise, because the total observed score is a weighted combination of the primary and secondary traits, unbalanced DIF will lead to confounding of impact with DIF (Camilli, 1992). There is a body of research on methods of purifying the matching score (for a recent study and further citations, see Zwick, 2012). This issue was not examined in the current study because if DIF effects and standard errors were estimated with similar accuracy regardless of the model underlying the DIF, the same items would tend to be identified for the purified matching score.

Another omitted condition was non-uniform DIF. In an MH or logistic-regression framework, non-uniform DIF occurs when the log-odds difference is not the same (not uniform) for all levels of the matching variable. If the data have a non-zero lower asymptote, the log-odds difference will be smaller for lower matching scores than for higher matching scores. However, in an IRT framework, non-uniform DIF is typically defined in terms of group-specific *a* or *c* parameters. Different *a*-parameters yield crossing DIF, a special case of non-uniform DIF where the DIF favors the reference group in one ability region and the focal group in another region (Li & Stout, 1996). The MH procedure used in this study can detect non-uniform DIF that does not cross (termed unidirectional DIF by Shealy & Stout, 1993) or DIF that crosses away from the center of the ability distribution (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). Because the parameter estimated in the MH procedure is the average log-odds difference, the MH procedure is not effective at detecting DIF that crosses near the center of the ability distribution (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993), unless the procedure is adapted to estimate the average log-odds difference above and below the crossing point separately (Mazor, Clauser, & Hambleton, 1994; Hidalgo, & López-Pina, 2004). In the MIRT model, one could model group-specific a_s. This would be equivalent to group differences in variance of θ_s. When one marginalizes over the secondary θ, this would yield a lower unidimensional *a*-parameter for the group with greater variance in θ_s.

Implications

It made virtually no difference in the Mantel-Haenszel Δ-difference estimates whether the DIF was modeled as a continuous secondary trait, using a MIRT model, or a difference in *b*-parameters using a unidimensional model. Thus, the previous DIF literature, which has largely simulated DIF with unidimensional models, should be applicable even

when researchers conceptualize DIF as a continuous trait that advantages some members and disadvantages other members of the same group. Simulating DIF with a unidimensional model still produces effect sizes that accurately reflect the average effect on the group. Additionally, there is no strong reason to prefer MIRT models to simulate DIF. Using unidimensional models with separate item parameters is simpler and easier to explain to most audiences.

The no impact/impact and 1PL/3PL conditions interacted in predictable ways: when there was a large group mean difference and the data were 3PL, easy DIF-free items appeared to favor the reference group and difficult DIF-free items appeared to favor the focal group. Although not new, this problem should not be forgotten.

Although only the Mantel-Haenszel DIF procedure was assessed in this study, results likely generalize to Δ-difference estimates from logistic regression (Monahan, McHorney, Stump, & Perkins, 2007; Zumbo, 2001). The SIBTEST effect size is on an entirely different metric, mean difference between the non-parametric ICCs, but there would be little reason to think that this difference would vary as long as the marginal ICCs from the MIRT model matched the unidimensional ICCs.

One path not explored in this study is the utility of MIRT models in recovering the secondary θ. The model could be expanded to include person characteristics leading to DIF, similar to IRT mixture models except that the secondary θ would be a continuous trait instead of a latent class. This would only be possible if multiple items were a function of the same secondary θ. Some might argue that this scenario no longer represents DIF but instead characterizes a meaningful, although perhaps unintended, secondary dimension. This phenomenon might be more interesting than DIF. Although from a policy standpoint the main concern may be the mean effect of DIF on demographic groups, researchers interested in explaining DIF would find it useful to model the effects on individuals.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting Sibtest detection procedure. *Behaviormetrika, 23*, 67-95.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*, 129-147.

Cohen A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133-148.

De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2*, 243-276.

De Boeck, P. (2008). Random IRT models. *Psychometrika, 73*, 533-559.

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

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

Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903-915.

Holland, P. W., & Thayer, D. T. (1985). *An alternate definition of the ETS delta scale of item difficulty* (Research Report RR-85-43). Princeton, NJ: Educational Testing Service. Available from ERIC database. (ED268148)

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

Jiang, H., & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics, 23*, 291-322.

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*, 647-677.

Li, Y., Brooks, G. P., Johanson, G. A. (2012). Item discrimination and type I error in the detection of differential item functioning. *Educational and Psychological Measurement, 72*, 847-861.

Mazor, K. M., Clauser, B. E., Hambleton, R. K. (1994). Identification of non-uniform DIF using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*, 284-291.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*, 92-109.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

R Core Team (2014). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org/.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361-373.

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

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*, 293-322.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF paradigm. *Applied Psychological Measurement, 20*, 355-371.

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

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

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Wainer, H., & Skorupski, W. P. (2005). Was it ethnic and social-class bias or statistical artifact? Logical and empirical evidence against Freedle's method for reestimating SAT scores. *Chance, 18*, 17-24.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. (2001, April). *Investigating DIF by the statistical modeling of the probability of endorsing an item: Logistic regression and extensions thereof.* Paper presented at the annual meeting of the National Council for Measurement in Education, Seattle, WA.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-197.

Zwick, R. (2012, May). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement (ETS RR-12-08)*. Princeton, NJ:ETS.

Appendix A

Parameters for items 13 - 60

Table A1:
Study 1: Parameters for DIF-free Items

	3PL		1PL		3PL		1PL
Item	*a*	*b*	*b*	Item	*a*	*b*	*b*
13	0.9	-2.139	-2.892	37	0.9	0.145	-0.301
14	1.1	-2.139	-3.082	38	1.1	0.145	-0.272
15	1.3	-2.139	-3.224	39	1.3	0.145	-0.252
16	1.5	-2.139	-3.332	40	1.5	0.145	-0.236
17	0.9	-1.530	-2.164	41	0.9	0.440	0.002
18	1.1	-1.530	-2.285	42	1.1	0.440	0.052
19	1.3	-1.530	-2.374	43	1.3	0.440	0.089
20	1.5	-1.530	-2.440	44	1.5	0.440	0.117
21	0.9	-1.104	-1.670	45	0.9	0.754	0.312
22	1.1	-1.104	-1.749	46	1.1	0.754	0.384
23	1.3	-1.104	-1.806	47	1.3	0.754	0.437
24	1.5	-1.104	-1.849	48	1.5	0.754	0.477
25	0.9	-0.745	-1.274	49	0.9	1.104	0.639
26	1.1	-0.745	-1.320	50	1.1	1.104	0.734
27	1.3	-0.745	-1.354	51	1.3	1.104	0.802
28	1.5	-0.745	-1.379	52	1.5	1.104	0.854
29	0.9	-0.440	-0.926	53	0.9	1.530	1.006
30	1.1	-0.440	-0.945	54	1.1	1.530	1.121
31	1.3	-0.440	-0.959	55	1.3	1.530	1.204
32	1.5	-0.440	-0.969	56	1.5	1.530	1.266
33	0.9	-0.145	-0.607	57	0.9	2.139	1.452
34	1.1	-0.145	-0.602	58	1.1	2.139	1.579
35	1.3	-0.145	-0.598	59	1.3	2.139	1.668
36	1.5	-0.145	-0.594	60	1.5	2.139	1.731

Note: The 3PL *a*'s are in the normal metric and would need to be multiplied by 1.7 if the logistic metric were preferred.

Table A2:
Study 2: Parameters for Additional 3PL DIF Items

	MIRT				unidimensional *b*-difference				
Item	a_1	a_2	MID	Δ-diff	a	b_{ref}	b_{foc}	c	Δ-diff
13	0.9	0.43	-2.14	-1.53	0.85	-2.53	-2.05	0.24	-1.52
24	1.5	0.50	-1.10	-1.49	1.33	-1.33	-0.99	0.20	-1.48
30	1.1	0.55	-0.44	-1.51	0.95	-0.74	-0.24	0.20	-1.50
38	1.1	0.65	0.15	-1.49	0.89	-0.14	0.45	0.19	-1.48
47	1.3	1.00	0.75	-1.49	0.87	0.55	1.32	0.19	-1.49
55	1.3	1.50	1.53	-1.04	0.70	1.73	2.88	0.19	-1.05
20	1.5	0.47	-1.53	-1.50	1.36	-1.75	-1.44	0.21	-1.48
25	0.9	0.50	-0.75	-1.51	0.80	-1.13	-0.57	0.20	-1.50
35	1.3	0.65	-0.15	-1.56	1.06	-0.42	0.08	0.20	-1.55
44	1.5	0.90	0.44	-1.55	1.07	0.20	0.80	0.19	-1.54
51	1.3	1.50	1.10	-1.48	0.68	1.08	2.23	0.19	-1.48
57	0.9	2.00	2.14	-0.58	0.43	3.86	6.12	0.20	-0.59
14	1.1	0.43	-2.14	1.52	1.03	-2.05	-2.44	0.23	1.50
23	1.3	0.50	-1.10	1.53	1.16	-0.98	-1.37	0.20	1.51
31	1.3	0.55	-0.44	1.47	1.12	-0.27	-0.69	0.20	1.45
37	0.9	0.65	0.15	1.54	0.73	0.53	-0.20	0.19	1.54
48	1.5	1.30	0.75	1.65	0.87	1.41	0.54	0.19	1.65
54	1.1	1.10	1.53	1.00	0.71	2.63	1.64	0.19	1.00
19	1.3	0.47	-1.53	1.53	1.18	-1.43	-1.79	0.21	1.51
26	1.1	0.53	-0.75	1.55	0.96	-0.59	-1.07	0.20	1.53
34	1.1	0.60	-0.15	1.51	0.92	0.10	-0.44	0.20	1.50
43	1.3	0.85	0.44	1.56	0.95	0.84	0.18	0.19	1.55
50	1.1	1.50	1.10	1.55	0.57	2.51	1.15	0.19	1.55
58	1.1	2.00	2.14	0.55	0.52	5.21	3.36	0.20	0.56

Note: In the MIRT model, all $c = 0.2$. DIF items 1-12 were the same as in Study 1 (Table 1). Other items not listed in this table were DIF-free and had the same parameters as in Study 1 (Table A1). The first 12 items listed were DIF items in both the 24-DIF and 36-DIF conditions in Study 2. The next 12 items were DIF items only in the 36-DIF condition. Sets of six items, separated by a horizontal line, were influenced by the same secondary trait in the MIRT model.

Table A3:
Study 2: Parameters for Additional 1PL DIF Items

	MIRT			unidimensional *b*-difference	
Item	a_2	MID	Δ-diff	b_{ref}	b_{foc}
13	0.70	-2.68	-1.57	-3.02	-2.35
24	0.70	-1.68	-1.53	-2.01	-1.36
30	0.70	-0.86	-1.50	-1.18	-0.54
38	0.70	-0.22	-1.50	-0.54	0.10
47	0.70	0.47	-1.50	0.15	0.79
55	0.45	1.32	-1.02	1.10	1.54
14	0.70	-2.82	1.57	-2.49	-3.16
23	0.70	-1.65	1.53	-1.32	-1.97
31	0.70	-0.87	1.50	-0.55	-1.19
37	0.70	-0.24	1.50	0.08	-0.55
48	0.70	0.51	1.50	0.83	0.19
54	0.45	1.24	1.02	1.46	1.02
20	0.70	-2.22	-1.55	-2.55	-1.89
25	0.70	-1.17	-1.51	-1.50	-0.85
35	0.70	-0.53	-1.50	-0.85	-0.21
44	0.70	0.14	-1.50	-0.17	0.46
51	0.70	0.87	-1.50	0.55	1.19
57	0.25	1.87	-0.58	1.74	1.99
19	0.70	-2.17	1.55	-1.84	-2.50
26	0.70	-1.21	1.51	-0.88	-1.53
34	0.70	-0.53	1.50	-0.21	-0.85
43	0.70	0.13	1.50	0.44	-0.19
50	0.70	0.86	1.50	1.18	0.54
58	0.25	1.86	0.58	1.99	1.74

Note: DIF items 1-12 were the same as in Study 1 (Table 2). Other items not listed in this table were DIF-free and had the same parameters as in Study 1 (Table A1). The first 12 items listed were DIF items in both the 24-DIF and 36-DIF conditions in Study 2. The next 12 items were DIF items only in the 36-DIF condition. Sets of six items, separated by a horizontal line, were influenced by the same secondary trait in the MIRT model.

Appendix B

Effects of impact and 3PL/1PL model independent of the form of the DIF

For the 3PL data, there was an interaction between impact and the item parameters. These interactions were the same for both the categorical and continuous DIF, and thus were not related to the research question. These interactions would be expected based on the literature on categorical DIF and are not new findings. However, a brief explanation may be interesting to those not familiar with this literature.

For the DIF-free items, when the impact was 1, the bias, although generally small, was negative (appeared to favor the reference group) for the easier, more discriminating items, but positive (appeared to favor the focal group) for the harder, less discriminating items. This was expected from previous research (Li, Brooks, & Johanson, 2012; Wainer & Skorupski, 2005; Zwick, 1990). If there is no correct guessing, more discriminating items tend to appear to favor the group that has the higher mean θ and less discriminating items appear to favor the other group. However, when there is some degree of correct guessing, the difference in average log-odds between the groups is narrowed for more difficult items, where many of the score levels show no difference between the groups because probability is at chance level for both groups, so in a relative sense these items favor the focal group.

For the DIF items,when there was no impact, the only small but perceptible bias was in items 2 and 8, the easiest, most discriminating items. Including the DIF studied item in the matching score creates a small group difference in the expected value of θ conditional on observed score such that the group favored by the item has a lower mean θ at any given matching score. This positively biases negative Δ-differences or negatively biases positive Δ-differences, but it is only noticeable for the easiest most discriminating items for the same reason that impact biased the Δ-differences most for the DIF-free items in this range.

Still considering the DIF items, when impact = 1, conditional on item discrimination, bias was more positive as item difficulty (MID) increased, as it was for the DIF-free items. The effect of item difficulty on bias was greater for more discriminating items. For the items that favored the reference group (negative true Δ-differences), increasingly positive bias meant that although the absolute value of the Δ-difference was overestimated for the easiest item, it was underestimated for the most difficult. For the items that favored the focal group (positive true Δ-differences), more positive bias meant that the absolute value of the Δ-difference was underestimated for the easiest item and overestimated for the most difficult. This was predictable from the results for the DIF-free items, which in turn were predictable from previous research. The bias was never large enough to change which group was favored by a given item.