Check for updates

Article

DIF Testing for Ordinal Items With Poly-SIBTEST, the Mantel and GMH Tests, and IRT-LR-DIF When the Latent Distribution Is Nonnormal for Both Groups

Applied Psychological Measurement
35(2) 145–164
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621610377450
http://apm.sagepub.com

Carol M. Woods[1]

Abstract

Differential item functioning (DIF) occurs when an item on a test, questionnaire, or interview has different measurement properties for one group of people versus another. One way to test items with ordinal response scales for DIF is likelihood ratio (LR) testing using item response theory (IRT), or IRT-LR-DIF. Despite the various advantages of IRT-LR-DIF, one disadvantage is that the latent variable is usually assumed to be normally distributed. If this normality assumption is violated, nonparametric alternatives such as the Mantel test, generalized Mantel–Haenszel (GMH) test, and poly-SIBTEST may be preferable. Simulations were carried out to compare IRT-LR-DIF to poly-SIBTEST and the GMH and Mantel tests when the latent density is nonnormal for both groups but presumed normal for IRT-LR-DIF. Results indicated that latent nonnormality detrimentally affected all three procedures, but IRT-LR-DIF was surprisingly more robust to latent nonnormality than all of the nonparametric approaches.

Keywords

differential item functioning, DIF, item bias, item response theory, ordinal items

Differential item functioning (DIF) occurs when an item on a questionnaire, test, interview, and so on has different measurement properties for one group of people versus another. Here, we deal with ordinal item responses modeled using Samejima's graded (1969, 1997) item response function (IRF). Items that function differently for the reference (R) and focal (F) groups may do so with respect to the discrimination or threshold parameters, or both. A differentially functioning (D-F) item may be more strongly associated with the primary construct for one group than the other, or the item may even measure a different construct for one group versus the other. At least one of the response options for a D-F item may be more easily endorsed by one group versus the other over some, or all, of the latent continuum.

[1]Washington University in St. Louis, Missouri, USA

Corresponding Author:
Carol M. Woods, Department of Psychology, University of Kansas, 1415 Jayhawk Blvd., Room 426, Lawrence, KS 66045-7556, USA
Email: carol.m.woods@gmail.com

Various approaches to DIF testing for items with ordinal response scales have been suggested. One method is likelihood ratio (LR) testing using item response theory (IRT), or IRT-LR-DIF (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993). In IRT-LR-DIF, nested two-group item response models with varying constraints are statistically compared to evaluate whether the response function for a particular item differs for the R and F groups. IRT-LR-DIF is advantageous because (a) measurement error is modeled; (b) separate item parameters are estimated for each group; (c) items with binary, ordinal, or nominal response scales may be tested; (d) both uniform and nonuniform DIF (defined by Camilli & Shepard, 1994, and Mellenbergh, 1989) may be evaluated; and (e) effect sizes are in readily comprehensible item–parameter units. Various software programs may be used, but analyses are particularly convenient when carried out using the IRTLRDIF v2.0b (Thissen, 2001) program.

One potential disadvantage of IRT-LR-DIF is the numerous assumptions. Typically, the IRFs are presumed to be logistic, and the distribution of the latent variable, θ, is presumed to be normal in both groups (though the mean and variance may differ for the R and F groups). If the assumption about the IRFs (Bolt, 2002), or the normality assumption about θ for one or both groups (Woods, 2008a), is incorrect, Type I error can be 2 to 4 times the nominal level. Woods (2008a) also observed overestimation of group differences in the item parameters and inaccurate estimation of the F-group mean and *SD* when normality was violated. These findings occur in the realistic case of a nonzero mean difference between groups.

Three alternative methods for DIF testing with ordinal items, the Mantel (1963) test (Zwick, Donoghue, & Grima, 1993), generalized Mantel–Haenszel test (GMH; Somes, 1986; Zwick et al., 1993), and poly-SIBTEST (SIB = simultaneous item bias; Chang, Mazzeo, & Roussos, 1996), require fewer assumptions than IRT-LR-DIF. When the assumptions made for IRT-LR-DIF are justifiable, these methods may be second choice because they have lower power, especially for nonuniform DIF, and matching is based on observed scores that can be a poor proxy for θ. However, when the assumption about the IRF or the shape of θ is violated, these other tests may be preferable to IRT-LR-DIF because they do not presume any particular IRF or distributional shape for θ. When the assumption about the IRF was incorrect (and there was a mean difference of 1 *SD*), poly-SIBTEST controlled Type I error better than IRT-LR-DIF (Bolt, 2002).

The purpose of the present study is to compare IRT-LR-DIF to poly-SIBTEST and the Mantel and GMH tests when the latent distribution is nonnormal for both groups, but presumed normal for IRT-LR-DIF. Although the performance of each of these procedures has been studied under various conditions, they have apparently never been compared in a single study for the case of latent nonnormality. Type I error and power of the DIF tests are evaluated as well as bias (and root mean square error [RMSE]) in the associated effect sizes. When the normality assumption about θ is violated (but the IRF is correctly specified), results are expected to be more accurate with poly-SIBTEST, the Mantel and GMH tests, and their associated effect sizes, than with IRT-LR-DIF and its associated effect size.

Another factor varied in the simulations was the pattern of DIF in the graded-model threshold parameters (b_{ij}s). Within an item, the DIF was either in the same direction and magnitude for all b_{ij}s (i.e., in a constant pattern) or not. Poly-SIBTEST and the Mantel test evaluate a net signed DIF effect (an overall effect considering that some responses may favor the R group whereas others favor the F group, within one item), and should do best under a constant DIF pattern and worst when exactly half of the DIF favors the R group and half favors the F group so that the apparent net effect is 0. Conversely, the GMH test evaluates a global, unsigned DIF effect over the response distribution and should do better when invariance effects are not constant. In IRT-LR-DIF, all b_{ij}s are considered together in one model comparison, but there is no reason to expect the test to perform any better with one pattern of DIF versus another.

IRT-LR-DIF

In IRT-LR-DIF, a series of two-group IRT models are fitted to data using Bock and Aitkin's (1981) scheme for marginal maximum likelihood. The mean and variance of θ are fixed to 0 and 1 (respectively) for the R group to identify the scale, and estimated for the F group. Anchor items, presumed to be DIF-free, link the metric of θ for the two groups. Item parameters for all anchors are constrained equal between groups in all models. Nonanchor items, called studied items, are tested individually for DIF.

For a particular studied item fitted with Samejima's graded (1969, 1997) model, the analysis begins with a general test for DIF in the discrimination parameter, a_i, the threshold parameters, b_{ij}s (j indexes thresholds), or both. The null (H_0) and alternative (H_a) hypotheses are as follows:

H_0: $a_{iF} = a_{iR}$ and $b_{ijF} = b_{ijR}$ for all j.
H_a: not all parameters for item i are group invariant.

A model with all parameters for the studied item constrained equal between groups is compared to a model with all parameters for the studied item permitted to vary between groups. The LR test statistic is –2 times the difference between the optimized log likelihoods, which is approximately χ^2-distributed, with *df* equal to the difference in free parameters. Statistical significance indicates the presence of DIF. If the general test is significant, follow-up tests are easily carried out to establish whether the DIF is due to unequal a_is or unequal b_{ij}s (or both). Because power and Type I error rates for the follow-up tests are highly dependent on those for the general test, only the general test is evaluated in the present study.

When assumptions are correct, Type I error for the general IRT-LR-DIF test is near the nominal level and the group-mean difference is recovered well under various realistic conditions (Ankenmann, Witt, & Dunbar, 1999; Bolt, 2002; Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998; Stark, Chernyshenko, & Drasgow, 2006; Sweeney, 1996; Wang & Yeh, 2003; Woods, 2009). Statistical power to detect uniform and nonuniform DIF increases with increases in sample size, item discrimination, the number of anchors, and the amount of DIF in the data (Ankenmann et al., 1999; Wang & Yeh, 2003; Woods, 2009). These results are based on item responses simulated from the two-parameter logistic, three-parameter logistic, or graded IRFs, with test lengths of 10, 15, 20, 25, 26, 30, 40, or 50 items, and group-mean differences of 0, .4, .5, or 1 *SD*. Sample sizes in these studies were equal for both groups ($N_R = N_F =$ 250, 300, 500, 1,000, or 2,000), or larger for the R group ($N_R/N_F =$ 1,000/300, 1,500/500, or 2,000/500). Anchors have been all other items, 1 item, or 10%, 16%, 20%, or 40% of the total.

When assumptions are violated, Type I error is inflated for IRT-LR-DIF. Simulations have revealed Type I error inflation[1] when the mean difference is nonzero and the assumption about the IRF or the θ distribution is incorrect (Bolt, 2002; Woods, 2008a), or when anchors function differently between groups (Stark et al., 2006; Wang & Yeh, 2003; Woods, 2009). If a researcher suspects or finds evidence that one of these assumptions is violated, results are likely to indicate that some invariant items function differently between groups. Bolt (2002) showed that poly-SIBTEST is preferable to IRT-LR-DIF when the assumption about the IRF is incorrect.

Average Unsigned Difference

An effect size associated with IRT-LR-DIF is the average unsigned difference (AUD) between the expected response functions (ERFs) for the F and R groups, weighted by the (presumed normal) F-group θ density (Wainer, 1993):

$$AUD = \frac{\sum_{q=1}^{Q} |ERF_R(\theta) - ERF_F(\theta)| g_F(\theta)}{Q}, \quad (1)$$

where $q = 1, 2, \ldots Q$ refers to quadrature points (Q is the total number of points). An ERF gives the item score (e.g., 1, 2, 3, 4, or 5) expected at each value of θ (Bolt, Hare, Vitale, & Newman, 2004; Steinberg & Thissen, 2006; Wainer, Sireci, & Thissen, 1991) and is computed as the sum of possible item scores (1, 2, 3, 4, or 5), each weighted by its expected probability according to the IRF. Here, θ is represented by 81 quadrature points between –4 and 4 in 0.1 increments, and the IRF is the graded model. For each studied item, the true AUD was computed using true item parameters for the ERF and the true F-group mean and *SD* (–0.5 and 1) for the normal $g_F(\theta)$, whereas the estimated AUD used estimated item parameters and estimates of the F-group mean and *SD* from the model that permitted the studied item's parameters to vary between groups.

The Mantel and GMH Tests

Holland and Thayer (1988) described how the Mantel–Haenszel χ^2 (MH; Mantel & Haenszel, 1959) could be used to test for DIF with items scored in two categories. This MH χ^2 can be extended for polytomous items in two ways, one that presumes the item response categories are ordered and compares means between groups (the Mantel test; Mantel, 1963) and another that does not presume ordering and compares the entire response distribution between groups (the GMH test; Somes, 1986). In one of the first simulations to evaluate these procedures in the context of DIF, Zwick et al. (1993) concluded that both the Mantel and GMH methods appear promising for ordinal items. However, the Mantel had greater power when the DIF was in a constant pattern for the b_{ij}s, whereas the GMH was more powerful when the DIF (for a single item) favored the R-group for half of the b_{ij}s and favored the F-group for the rest of the b_{ij}s (Zwick et al., 1993).

Both the GMH and the Mantel indicate the relation between the item and group membership, controlling for a matching score, which is typically an observed sum of item scores (i.e., a summed score). Primarily uniform (not nonuniform) DIF is detected. For best control of Type I error, the matching score should be based on DIF-free items plus the studied item (Su & Wang, 2005; Wang & Su, 2004; Zwick et al., 1993).

The GMH statistic may be expressed as

$$\chi^2_{GMH} = \left[\sum_{m=o}^{T} \mathbf{N}_m - \sum_{m=o}^{T} E(\mathbf{N}_m)\right]' \left[\sum_{m=o}^{T} V(\mathbf{N}_m)\right]^{-1} \left[\sum_{m=o}^{T} \mathbf{N}_m - \sum_{m=o}^{T} E(\mathbf{N}_m)\right], \quad (2)$$

where m is a score on the matching criterion ($m = 0, 1, 2, \ldots, T$), and $\mathbf{N}_m$ is a vector of $c - 1$ R-group frequencies (c = total number of categories). $E(\,)$ is the expected value under the null hypothesis of no DIF and $V(\,)$ is the variance; both are functions of marginal frequencies (see e.g., Zwick et al., 1993, p. 239, for details). Under the null hypothesis, the GMH statistic is distributed as $\chi^2(df = c - 1)$.

The Mantel statistic is given by

$$\chi^2_{Mantel} = \frac{\left(\sum_{m=0}^{T} F_m - \sum_{m=0}^{T} E(F_m)\right)^2}{\sum_{m=0}^{T} \sigma^2_{F_m}}, \quad (3)$$

where F_m is the sum of scores for the F group at the mth level of the matching criterion and σ^2 is the variance. Under the null hypothesis, the Mantel statistic is distributed as χ^2 ($df = 1$).

Power depends on the pattern of DIF in the b_{ij}s (i.e., the number of b_{ij}s that differ between groups, and in what directions), the sample size, degree of item discrimination, and amount of DIF, and Type I error can be well controlled under certain circumstances (Ankenmann et al., 1999; Chang et al., 1996; Penfield, 2007; Penfield & Algina, 2003; Su & Wang, 2005; Wang & Su, 2004; Welch & Hoover, 1993; Zwick et al., 1993; Zwick, Thayer, & Mazzeo, 1997). The Mantel test shows Type I error inflation under two realistic conditions—item-varying discrimination and a nonzero mean difference (Ankenmann et al., 1999; Chang et al., 1996; Su & Wang, 2005; Wang & Su, 2004; Zwick et al., 1993; Zwick et al., 1997). Increasing the number of items used for the matching score can help reduce Type I error (Wang & Su, 2004), probably because of improved reliability.

Simulations described above used ordinal item responses generated from the graded (Samejima, 1969, 1997), partial credit (Masters, 1982), or generalized partial credit (Muraki, 1992) IRFs, group-mean differences of 0, 0.5, 1, or 1.5 *SD*, and samples sizes that were equal for both groups ($N_R = N_F =$ 250, 500, 1,000, 2,000) or greater for the R group ($N_R/N_F =$ 1,500/500 or 2,000/500). Matching was based on summed scores (including the studied item except in the study by Welch & Hoover, 1993) composed of 10, 20, 25, 26, 30, 25, 50, or 71 binary or ordinal items.

Liu-Agresti Cumulative Common Odds Ratio Estimator

An effect size associated with the Mantel test is the Liu-Agresti cumulative common odds ratio estimator ($\hat{\varphi}_{LA}$; Liu & Agresti, 1996) described in the context of DIF by Penfield and Algina (2003). The estimator is essentially a sum of all possible odds ratios, where the odds ratios are computed for individual group by item (2×2) tables constructed by collapsing item response categories at all possible places that preserve ordering of the item response. The odds are, for instance, of a higher versus lower response for the R group versus F group, and the sum is over all levels of the matching criterion. It is assumed for computation of $\hat{\varphi}_{LA}$ that the response categories are ordered and that the cumulative log odds ratio is constant over categories.

The estimator is

$$\hat{\varphi}_{LA} = \frac{\sum_{m=0}^{T} \frac{1}{N_m} \sum_{j=1}^{c-1} A_{mj} D_{mj}}{\sum_{m=0}^{T} \frac{1}{N_m} \sum_{j=1}^{c-1} B_{mj} C_{mj}}, \tag{4}$$

where m is a score on the matching criterion, N_m is the total number of people (R + F group members) at that score level, and A_{mj}, B_{mj}, C_{mj}, and D_{mj} refer to frequencies in a group by item response (2×2) table constructed by collapsing response categories such that the response is either $\leq j$ or $> j$. For better interpretation, Penfield and Algina (2003) take the natural log of the inverse of $\hat{\varphi}_{LA}$, and $\log(\frac{1}{\hat{\varphi}_{LA}}) = 0$ suggests no DIF. The DIFAS program (Penfield, 2005) is freely available from Penfield and provides an easy way to compute $\log(\frac{1}{\hat{\varphi}_{LA}})$.

When the DIF is constant over b_{ij}s, the true LA effect size, φ_{LA}, is equal to $a_i\gamma$, where $a_i =$ true discrimination and $\gamma =$ true amount of DIF in each b_{ij}. This definition of the true LA effect size holds only for the case of equal a_i parameters between groups and is therefore used to evaluate bias of the effect size estimates only for items without group differences in a_is.

In the present study, $\log(\frac{1}{\varphi_{LA}})$ is computed for simulation conditions with constant DIF in the b_{ij}s, and compared to the true LA effect size, $\log(\frac{1}{\varphi_{LA}}) = a_i\gamma$. In previous simulations, the mean (over replications) $\log(\frac{1}{\varphi_{LA}})$ was somewhat biased when $\gamma \neq 0$, especially when there was a group mean difference on θ (Penfield & Algina, 2003). These authors also noted that φ_{LA} and the mean $\hat{\varphi}_{LA}$ are expected to differ whenever a_i varies over items because the matching criterion is θ for φ_{LA} versus a summed score for $\hat{\varphi}_{LA}$.

Poly-SIBTEST

Shealy and Stout (1993) introduced SIBTEST to detect uniform DIF in binary items, which extends and improves on the standardization approach (Dorans & Holland, 1993). With minor modifications, SIBTEST was further developed for items with ordinal responses (poly-SIBTEST; Chang et al., 1996). Poly-SIBTEST is considered nonparametric because few assumptions are required. A subset of items must be presumed approximately unidimensional and DIF-free, and the R and F groups must be independent, but no particular IRF or distribution for θ is assumed. Studied items may be assessed in combination when differential test functioning is of interest, or one at a time when DIF is of interest. Here, we focus on testing each item individually for DIF.

The poly-SIBTEST effect size estimator is $\hat{\beta} = \sum_{m=0}^{T} w_m(\bar{Y}_{Rm} - \bar{Y}_{Fm})$, where m is a score on the matching criterion ($m = 0, 1, 2, \ldots, T$), $\bar{Y}_{Rm}$ and $\bar{Y}_{Fm}$ are mean item scores for the R or F group with score m, and w_m is a weight equal to the proportion of all examinees with score m. The matching criterion is the sum of scores on items presumed DIF-free. In the present simulations, bias in the poly-SIBTEST estimator is evaluated with the true AUD designated as the true effect size, in accordance with the definition given by Shealy and Stout (1993, p. 167). Some bias is expected because the matching criterion is θ for AUD versus a summed score for $\hat{\beta}$.

The mean item scores, $\bar{Y}_{Rm}$ and $\bar{Y}_{Fm}$, are adjusted for group mean differences on the latent variable using a correction that involves the regression of true score on observed score (Chang et al., 1996; Shealy & Stout, 1993). The slope of the regression is the reliability of the test, excluding the studied item, which is equal to Cronbach's alpha with a guessing correction (Zwick et al., 1997). The currently available version (1.7) of poly-SIBTEST software, which was used for the present simulations, assumes that the regression of true score on observed score is linear. For this assumption to hold, there must be enough anchor items. Based on simulations with binary data, Shealy and Stout (1993) suggested at least 20 anchors for SIBTEST (p. 170); it is unclear what the recommendation is for poly-SIBTEST.

The effect size, $\hat{\beta}$, is the expected amount of DIF for a randomly selected F-group member. With sufficient sample size, $\hat{\beta}_{UNI} = \frac{\hat{\beta}}{\hat{\sigma}_{\beta}}$ is normally distributed with $\hat{\sigma}_{\beta} = \sqrt{\hat{\sigma}^2_{\beta}}$ and

$$\hat{\sigma}^2_{\beta} = \sum_{m=0}^{T} w_m^2 \left(\frac{\hat{\sigma}^2_{Rm}}{N_{Rm}} + \frac{\hat{\sigma}^2_{Fm}}{N_{Fm}} \right), \tag{5}$$

where $\hat{\sigma}^2_{Rm}$ is the sample variance of the studied-item scores for R-group examinees with score m and N_{Rm} is the total number of R-group examinees with score m ($\hat{\sigma}^2_{Fm}$ and N_{Fm} are defined analogously for the F group). With $\alpha = .05$, there is evidence of significant DIF if $|\hat{\beta}_{UNI}| > 1.96$.

In simulations with item-varying discrimination and a nonzero mean difference (Bolt, 2002; Chang et al., 1996; Zwick et al., 1997), Type I error has been near nominal or moderately inflated for poly-SIBTEST and more accurate for poly-SIBTEST than for the Mantel test. Statistical

power to detect uniform DIF increased with increases in sample size, effect size, or item discrimination. In these studies, ordinal item responses were generated from the graded, partial credit, generalized partial credit, or 2 parameter sequential response (Mellenbergh, 1995) IRF, and matching scores included 20, 25, 29, 30, or 50 binary or ordinal items. The mean difference was 0 or 1 and the sample sizes were equal for both groups ($N_R = N_F =$ 300, 500, 1,000, or 2,000).

Method

Design

There were 72 independent conditions and 300 replications per condition. Manipulated variables were the total sample size ($N =$ 2,000 or 1,000), number of items ($k =$ 12, 24, or 48), proportion of differentially functioning (D-F) items[2] (1/6 or 1/3), pattern of DIF in the b_{ij}s (constant or empirically observed, explained below), and latent distributional shapes (three patterns, explained below). The F:R ratio of sample sizes was 2:3; thus, variations were 800:1,200 and 400:600. The proportions of D-F items used here were chosen in accordance with those observed in 18 applications of IRT-LR-DIF (cited in Woods, 2009).

In each condition, some items were DIF-free designated anchors and some were studied items (i.e., tested for DIF). One third of the studied items were DIF-free so that Type I error could be evaluated. Another third of the studied items were group variant in a_i and at least one b_{ij}, and the last third of studied items were group variant in at least one b_{ij} (but not in a_i). As an example, with $k =$ 24 and 1/3 D-F items, there were 12 anchors and 12 studied items: 4 with unequal-a DIF, 4 with equal-a DIF, and 4 that were DIF-free. With $k =$ 24 and 1/6 D-F items, there were 18 anchors and 6 studied items: 2 with unequal-a DIF, 2 with equal-a DIF, and 2 that were DIF-free.

The two types of simulated DIF are analogous to nonuniform and uniform DIF, respectively, but those terms may not be exactly accurate for polytomous items because nonuniform DIF is not necessarily a function of only a_i. The terms ''equal-a'' and ''unequal-a'' are used instead. Unequal-a items were simulated to be more discriminating or more highly related to the primary latent construct for the R versus F group.

Constant Versus Empirically Observed DIF in b_{ij}

There were two patterns of DIF simulated for the b_{ij}s: constant and empirically observed. For the constant pattern, F-group thresholds were defined as $b_{i1F} = b_{i1R} + \gamma$, $b_{i2F} = b_{i2R} + \gamma$, $b_{i3F} = b_{i3R} + \gamma$, and $b_{i4F} = b_{i4R} + \gamma$, where γ was equal to one of five equally likely values (.3, .4, .5, .6, or .7). A random number from a uniform distribution determined γ for a given item. All DIF favored the R group and was held constant over thresholds (i.e., $b_{ijF} > b_{ijR}$ for all j). The constant pattern indicates that the R group is more likely than the F group to endorse each response option, and that the group difference is the same for each response. It is useful to include this pattern for comparison to other simulation studies (because it is commonly simulated), the Mantel and poly-SIBTEST should perform best with this pattern, and because it permits evaluation of φ_{LA}, but it is unclear how frequently the pattern occurs in real data.

The empirically observed set of DIF patterns provides greater external validity. Here, the pattern of DIF in b_{ij}s was selected, based on the magnitude of the randomly drawn b_{i1R} (i.e., the first R-group b_{ij}) from among nine patterns observed in applications of IRT-LR-DIF (Orlando & Marshall, 2002; Reise, Widaman, & Pugh, 1993). The patterns and criteria used for selecting them are listed in Table 1. When an unordered set of b_{ijF}s (i.e., a set for which this was untrue:

Table 1. Empirically Observed Patterns of DIF in Simulated b_{ij}

Application				Adjustment to b_{ijR} to get b_{ijF}			
Study	Item	$\hat{b}_{i1R}$	Simulation criterion	b_{i1}	b_{i2}	b_{i3}	b_{i4}
Orlando	11	0.73	b_{i1R}s $\geq$ 0.73	–0.67	–0.27	–0.01	+0.53
Orlando	8	0.72	0.16 $\leq b_{i1R} <$ 0.73	–1.16	–0.53	–0.19	+0.19
Orlando	7	0.16	0.09 $\leq b_{i1R} <$ 0.16	–0.95	–0.02	+0.20	+0.89
Orlando	2	0.09	–0.44 $\leq b_{i1R} <$ 0.09	–1.06	–0.32	+0.06	+0.59
Reise	3	–0.44	–0.60 $\leq b_{i1R} <$ –0.44	+0.54	+0.09	–0.52	–0.81
Reise	5	–0.60	–0.83 $\leq b_{i1R} <$ –0.60	+0.16	–0.12	–0.34	–0.83
Orlando	4	–0.83	–1.03 $\leq b_{i1R} <$ –0.83	–0.01	+0.60	+0.43	+0.97
Reise	2	–1.03	–1.14 $\leq b_{i1R} <$ –1.03	+0.30	+0.25	+0.09	+0.03
Orlando	16	–1.14	b_{i1R}s < –1.14	–0.12	–0.29	–0.04	+1.52

Note: DIF = differential item functioning; Orlando = Orlando & Marshall (2002); Reise = Reise, Widaman, & Pugh (1993); Item = item number in the application; subscript F = focal group; subscript *R* = reference group; subscript *i* = indexes items; $\hat{b}_{i1R}$ = threshold parameter for the first category estimated in the empirical study; b_{i1R} = threshold parameter for the first category randomly drawn from a certain distribution in the simulation study.

$b_{i1F} < b_{i2F} < b_{i3F} < b_{i4F}$) was produced for item *i*, the C++ program rejected it and regenerated true b_{ij}s (both R and F) for item *i* until an ordered set of b_{ijF}s was obtained.

Latent Distributional Shapes

The latent distributions for the R and F groups, $g_R(\theta)$ and $g_F(\theta)$, were either both normal, or both positively skewed and leptokurtic. Two different nonnormal curves were used; these are plotted in Figure 1. Each curve is a mixture of two normals with skewness and kurtosis coefficients near to those estimated for real data using IRT, with the latent distribution estimated nonparametrically using B-splines (Woods, 2006; Woods & Thissen, 2006). Parameters of the mixtures are as follows, with μ = mean, σ = standard deviation, and *mp* = mixing proportion: Skew-1: $\mu_1 = -0.50$, $\mu_2 = 0.691$, $\sigma_1 = 0.40$, $\sigma_2 = 1.157$, $mp_1 = 0.58$, $mp_2 = 0.42$, and Skew-2: $\mu_1 = -0.253$, $\mu_2 = 2.192$, $\sigma_1 = 0.609$, $\sigma_2 = 1.045$, $mp_1 = 0.897$, $mp_2 = 0.103$. Regardless of the shapes of the distributions, the R-group mean and *SD* were 0 and 1 (respectively) and the F-group mean and *SD* were –0.5 and 1 (respectively).

Implementation

A C++ program was written to generate data, perform IRT-LR-DIF, compute the AUD and $\log(\frac{1}{\hat{\varphi}_{LA}})$, run the SAS freq procedure to compute the Mantel and GMH tests, and operate the poly-SIBTEST program (version 1.7; Roussos & Stout, 2005). For all methods, exactly the same item responses were used and significance tests were based on $\alpha = .05$. Poly-SIBTEST was run with w_m = proportion of all examinees with score *m* and guessing parameter = 0 (guessing was not applicable because the simulated item responses were Likert-type). Data for a particular *m* were excluded from the calculation of $\hat{\beta}$ if there were fewer than two observations per cell at that score level. For poly-SIBTEST, the matching criterion was the sum of anchor-item scores. For the Mantel and GMH tests, the matching criterion was the sum of anchor-item scores plus the studied-item score.

For IRT-LR-DIF, slightly modified source code from Thissen's IRTLRDIF (version 2.0b, 2001) software was used, with no changes to the estimation procedures. Bock and Aitkin's (1981) EM-MML estimation scheme was used to fit Samejima's graded model to the data

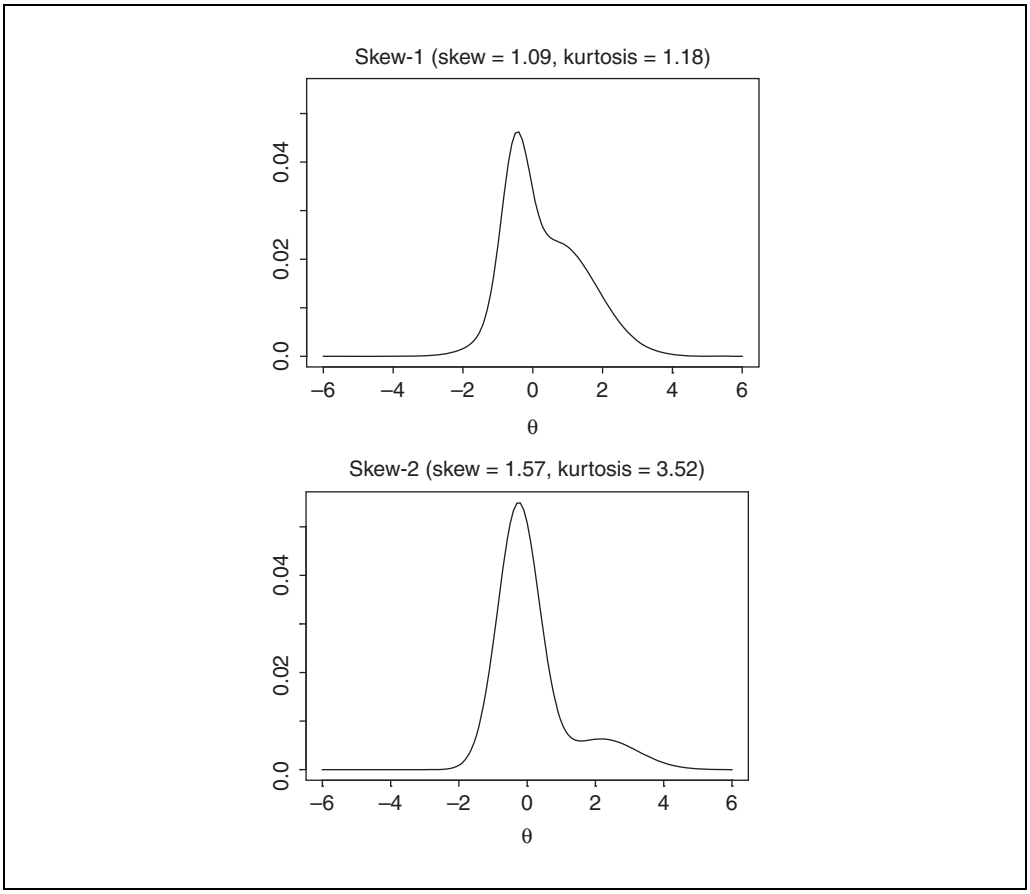

Figure 1. Nonnormal curves used as true latent distributions in the simulations (three has been subtracted from kurtosis)

with normal distributions assumed for $g_R(\theta)$ and $g_F(\theta)$. The mean and variance of $g_R(\theta)$ were fixed at 0 and 1 whereas the mean and variance of $g_F(\theta)$ were estimated. The latent variable was represented with rectangular quadrature, ranging from –4 to 4 in increments of 0.1 (81 points). The maximum number of EM cycles was 1,000 for fittings with the parameters for studied item *i* constrained equal in both groups, and 500 for fittings with the parameters for studied item *i* permitted to vary between groups. A fitting was declared converged when the parameter that was changing the most between EM cycles changed less than .0001.

Data

Five-category ordinal item response data were generated from the graded model (without the scaling constant 1.701). R-group item parameters were randomly drawn from certain distributions separately for each replication, and F-group parameters were defined in relation to them. These distributions (including truncation limits given below) were selected based on an examination of item parameters estimated from various educational and psychological scales (Hill, 2004), in combination with information obtained from a review of 18 published applications of IRT-LR-DIF (cited in Woods, 2009).

R-group discrimination parameters (a_{iR}) were drawn from $N(\mu = 1.7, \sigma = 0.6)$ with truncation on the upper end at 4.0, and on the lower end at 0.8. Truncation at 4.0 prevented a_{iR} from becoming unrealistically large, and the maximum amount of DIF in a_i was 0.7, so truncation at 0.8 ensured that a_{iF} was never less than 0.1. The first R-group threshold, b_{i1R}, was drawn from $N(\mu = -0.4, \sigma = 0.9)$ with truncation at –2.5 and 1.5. Subsequent thresholds were created by adding a randomly drawn value, d_{ihR}, to the immediately previous threshold (h counts differences between consecutive b_{ijR}s, where $j = 1, 2, 3, 4$). The difference between adjacent b_{ijR}s was drawn from $N(\mu = 0.9, \sigma = 0.4)$, with truncation at 0.1 and 1.5.

The F-group discrimination parameter (a_{iF}) was defined as $a_{iR} - \delta$, with δ equal to either 0 (for items without unequal-a DIF), or to one of five equally likely values (.3, .4, .5, .6, or .7). A random number from a uniform distribution determined δ for a given item. F-group thresholds were either equal to R-group thresholds (for DIF-free items), or defined with either constant or empirically observed DIF as described above.

When an item happened to be simulated with a 0 cell frequency for either the R or F group (i.e., 0 simulees responded in one or more of the five categories), categories for the item were collapsed for both groups. The mean number of collapses, divided by the total number of items, for each data set (for each simulation condition), was recorded.

Outcomes

Type I error rate, statistical power, and absolute bias and RMSE in effect size estimates were examined. For each condition, Type I error was computed by averaging (over replications) the proportion of DIF-free studied items with a significant test ($\alpha = .05$, two sided). Binomial confidence intervals were used to help interpret how much variability around the nominal error rate is acceptable given the number of replications. The 95% error limits are: $\alpha \pm 1.96\left(\sqrt{\frac{\alpha(1-\alpha)}{nreps}}\right) = .03$ to .07, where $nreps = 300$ and $\alpha = .05$.

Statistical power was calculated, separately for items with unequal-a versus equal-a DIF, as the average proportion of D-F items with significant tests. Values given can be interpreted as the percentage or proportion of items with DIF that were detected; thus, 100 (or 1) minus that value is the percentage (or proportion) of items with DIF that were not detected.

Absolute bias and RMSE were computed for the effect sizes, separately for studied items with unequal-a, equal-a, and no DIF.

Simulation Results

Category Collapses

The number of response categories collapsed as a result of a 0 cell frequency was greater with the smaller sample size and the normal versus nonnormal latent distributions, but there were no consistent differences in collapses between the constant and empirical DIF patterns. Differences due to test length and proportion of D-F items disappeared when the mean number of collapses was divided by the total number of items on the test. The percentage of items that required collapsing was approximately 5% (nonnormal θ) or 11% (normal θ) with $N = 1{,}000$, and 2% (nonnormal θ) or 6% (normal θ) with $N = 2{,}000$.

Type I Error

Type I error rates are shown in Figures 2 and 3 for the constant and empirical DIF patterns, respectively. In both figures, results are on the left for normal g_R and g_F, in the middle for

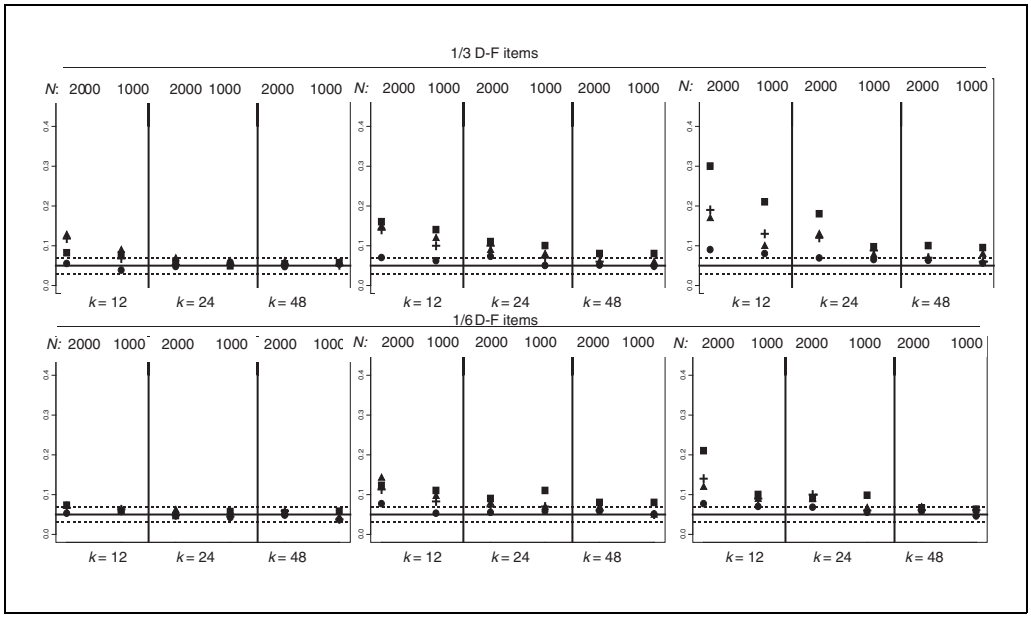

Figure 2. Type I error rates ($\alpha = .05$) for the constant DIF pattern

Note: Latent densities are normal g_R and g_F (left), Skew-1 g_R and Skew-2 g_F (middle), or Skew-2 g_R and Skew-1 g_F (right) and the fraction of D-F items is 1/3 (upper) or 1/6 (lower). Dashed lines mark the error limits at .03 and .07; k = number of items. For each k, results are given for descending N: 2,000 (left) and 1,000 (right). • = IRT-LR-DIF; ■ = poly-SIBTEST; ▲ = Mantel test; + = GMH test.

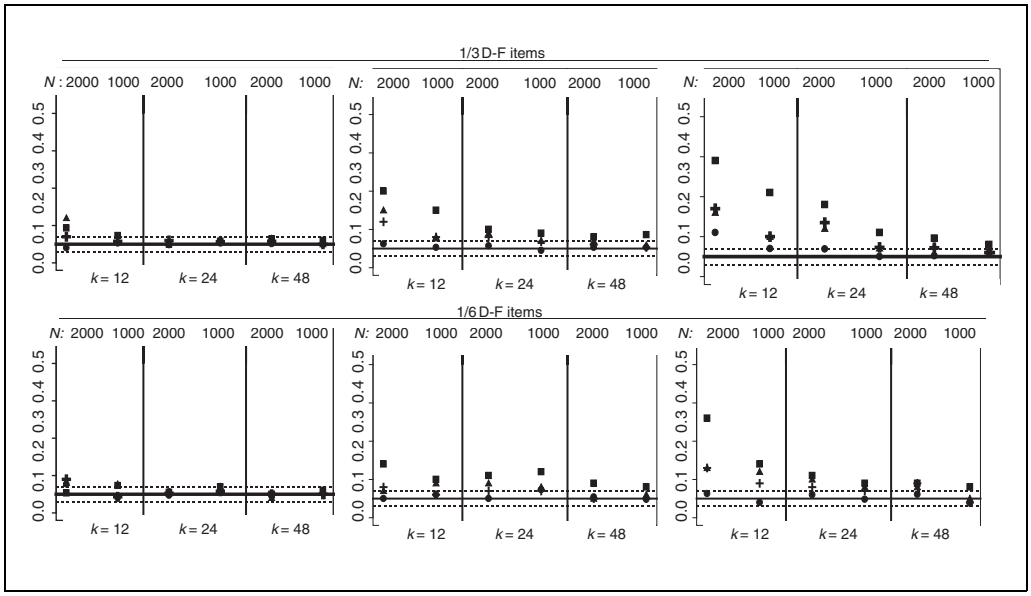

Figure 3. Type I error rates ($\alpha = .05$) for the empirical DIF pattern

Note: Latent densities are normal g_R and g_F (left), Skew-1 g_R and Skew-2 g_F (middle), or Skew-2 g_R and Skew-1 g_F (right) and the fraction of D-F items is 1/3 (upper) or 1/6 (lower). Dashed lines mark the error limits at .03 and .07; k = number of items. For each k, results are given for descending N: 2,000 (left) and 1,000 (right). • = IRT-LR-DIF; ■ = poly-SIBTEST; ▲ = Mantel test; + = GMH test.

Skew-1 g_R and Skew-2 g_F, and on the right for Skew-2 g_R and Skew-1 g_F. For each scale length, rates are given for decreasing N (left: 2,000, right: 1,000). Dashed lines mark the 95% error limits. Error was closer to the nominal level as k increased, probably because of the increase in anchors. The more anchors, the more reliable the matching criterion. Error was higher as the proportion of D-F items increased, probably because the number of anchors decreased.

Constant DIF pattern with normal g_R and g_F. When the latent distributions were normal for both groups (left, Figure 2), Type I error for IRT-LR-DIF was within the error limits in all conditions. In contrast, error rates for poly-SIBTEST and the Mantel and GMH tests were within the limits only when the matching criterion was based on at least about 12 items. Error was within the limits for all methods when the proportion of D-F items was 1/6, or for higher proportions if $k = 48$.

Constant DIF pattern with nonnormal g_R and g_F. When the latent distributions were nonnormal (middle and right, Figure 2), Type I error for IRT-LR-DIF was no longer consistently within the error limits. Inflation was worse for shorter scales (fewer anchors) and when the R-group distribution was Skew-2 instead of Skew-1. Contrary to prediction, error was consistently closer to the nominal level for IRT-LR-DIF versus poly-SIBTEST and the GMH and Mantel tests. Poly-SIBTEST was usually the least accurate, but the direction of difference between the Mantel and GMH tests was not consistent. When 1/6 of items functioned differentialy and N and k were larger, Type I error was within the limits for all methods except poly-SIBTEST.

Empirical DIF pattern with normal g_R and g_F. With the empirical DIF pattern and normality for both distributions (left, Figure 3), Type I error was within the limits for all methods if there were at least 12 anchors. With fewer anchors, IRT-LR-DIF still usually performed well, but error rates for the other methods were higher.

Constant DIF pattern with nonnormal g_R and g_F. With the empirical DIF pattern and nonnormal latent distributions (middle and right, Figure 3), Type I error was again inflated most severely for poly-SIBTEST and least severely for IRT-LR-DIF, with the Mantel and GMH tests in between. There were many conditions for which error was within the limits around the nominal level for IRT-LR-DIF, but there were no conditions for which this was true for poly-SIBTEST.

Power

Statistical power for detecting equal-a and unequal-a DIF with $N = 1{,}000$ is listed in Table 2 for 1/3 D-F items and Table 3 for 1/6 D-F items. Results for $N = 2{,}000$ were similar except that power was slightly greater (complete results are available upon request from the author). Shading indicates that Type I error was outside the limits for that condition. For IRT-LR-DIF and the GMH test, power for both equal-a and unequal-a DIF was generally high, and tended to be higher for the empirical versus constant DIF pattern. For the Mantel test and poly-SIBTEST, power was high for equal-a DIF and moderate for unequal-a DIF with the constant pattern, but power for equal-a versus unequal-a DIF was similar (and somewhat lower) for the empirical pattern.

Bias and RMSE in Effect Sizes

Absolute bias and RMSE of the effect sizes for studied items with unequal-a, equal-a, or no DIF is given in Tables 4 (constant DIF pattern) and 5 (empirical DIF pattern) for $N = 2{,}000$. Results were similar for $N = 1{,}000$ (and available on request from the author). Bias and RMSE were always near 0 for the AUD; thus, violating the assumption of normality for the F-group density did not appear to cause a problem. Although the RMSE for φ_{LA} was often rather large, this was primarily due to variability rather than bias: Absolute bias for φ_{LA} was always low. For β, RMSE was moderate to high except in one case (items without DIF; $k = 12$, 1/6 D-F items). Bias for β

Table 2. Statistical Power; $N_R = 600$, $N_F = 400$; 1/3 D-F Items

	Constant DIF in b_{ij}						Empirically observed DIF in b_{ij}					
	$k = 12$		$k = 24$		$k = 48$		$k = 12$		$k = 24$		$k = 48$	
	a&*b*	*b*	*a*&*b*	*b*	*a*&*b*	*b*	*a*&*b*	*b*	*a*&*b*	*b*	*a*&*b*	*b*
Normal g_R and g_F												
IRT-LR	93	90	94	90	95	91	100	99	100	98	100	99
Poly-SIB	68	92	70	95	68	94	77	74	76	74	77	74
Mantel	72	93	71	94	69	94	76	75	76	75	79	76
GMH	90	89	89	89	88	90	100	98	100	98	100	98
Skew-1 g_R, Skew-2 g_F												
IRT-LR	89	88	91	89	93	89	100	99	100	99	100	99
Poly-SIB	68	92	69	94	62	94	75	72	77	73	75	70
Mantel	68	91	67	93	65	92	74	70	76	71	76	71
GMH	82	86	83	89	83	86	100	99	100	99	99	99
Skew-2 g_R, Skew-1 g_F												
IRT-LR	91	88	91	88	94	91	100	99	100	99	100	99
Poly-SIB	74	97	67	95	65	93	75	70	74	74	72	70
Mantel	69	91	66	90	67	92	76	69	75	76	74	71
GMH	86	86	84	85	82	87	100	99	100	99	99	99

Note: Shading indicates that Type I error was inflated; k = number of items, D-F = differentially functioning; *a*&*b* or *b* refers to the item parameters that differed between groups; g_R and g_F = latent distributions for reference and focal groups; D-F = differentially functioning; DIF = differential item functioning; IRT = item response theory; LR = likelihood ratio; SIB = simultaneous item bias; GMH = generalized Mantel–Haenszel.

Table 3. Statistical Power; $N_R = 600$, $N_F = 400$; 1/6 D-F Items

	Constant DIF in b_{ij}						Empirically observed pattern of DIF in b_{ij}					
	$k = 12$		$k = 24$		$k = 48$		$k = 12$		$k = 24$		$k = 48$	
	a&*b*	*b*	*a*&*b*	*b*	*a*&*b*	*b*	*a*&*b*	*b*	*a*&*b*	*b*	*a*&*b*	*b*
Normal g_R and g_F												
IRT-LR	94	89	94	91	96	90	100	99	100	99	100	98
poly-SIB	71	92	66	96	66	93	77	75	78	74	76	74
Mantel	73	91	67	95	68	87	79	76	80	74	78	76
GMH	88	86	87	90	93	89	100	98	100	99	100	98
Skew-1 g_R, Skew-2 g_F												
IRT-LR	92	86	90	91	94	90	100	98	100	99	100	99
Poly-SIB	70	93	67	96	62	92	79	71	73	72	72	73
Mantel	69	91	67	94	63	93	77	70	75	75	75	73
GMH	86	85	80	90	80	87	100	98	99	98	99	99
Skew-2 g_R, Skew-1 g_F												
IRT-LR	95	87	93	92	94	91	100	100	100	99	100	99
Poly-SIB	71	96	69	95	61	90	76	70	73	74	73	71
Mantel	68	91	67	92	64	91	78	70	75	74	74	71
GMH	87	86	85	87	81	85	100	100	99	98	99	98

Note: Shading indicates that Type I error was inflated; k = number of items, D-F = differentially functioning; *a*&*b* or *b* refers to the item parameters that differed between groups; g_R and g_F = latent distributions for reference and focal groups; D-F = differentially functioning; DIF = differential item functioning; IRT = item response theory; LR = likelihood ratio; SIB = simultaneous item bias; GMH = generalized Mantel–Haenszel.

Table 4. Absolute Bias (and RMSE) in Effect Sizes; $N_R = 1{,}200$, $N_F = 800$; Constant DIF in b_{ij}

	$k = 12$		$k = 24$		$k = 48$	
Proportion of D-F items	1/6	1/3	1/6	1/3	1/6	1/3
Normal g_R and g_F						
Unequal-a: AUD	.00 (.01)	.00 (.01)	.00 (.01)	.00 (.01)	.00 (.01)	.00 (.01)
β	.12 (.23)	.12 (.26)	.11 (.24)	.12 (.24)	.10 (.24)	.11 (.25)
Equal-a: AUD	.00 (.01)	.00 (.01)	.00 (.01)	.00 (.01)	.00 (.01)	.00 (.01)
φ_{LA}	.03 (.14)	.04 (.58)	.02 (.52)	.04 (.56)	.01 (.55)	.01 (.54)
β	.23 (.25)	.24 (.24)	.23 (.24)	.23 (.23)	.23 (.24)	.23 (.24)
No DIF: AUD	.01 (.02)	.02 (.01)	.01 (.01)	.01 (.01)	.01 (.01)	.01 (.01)
φ_{LA}	.01 (.13)	.02 (.15)	.00 (.13)	.01 (.53)	.01 (.61)	.00 (.56)
β	.01 (.05)	.01 (.27)	.00 (.25)	.00 (.24)	.00 (.27)	.00 (.26)
Skew-1 g_R, Skew-2 g_F						
Unequal-a: AUD	.00 (.02)	.00 (.02)	.00 (.01)	.00 (.02)	.00 (.01)	.00 (.01)
β	.12 (.25)	.13 (.25)	.11 (.24)	.11 (.26)	.08 (.25)	.10 (.22)
Equal-a: AUD	.00 (.02)	.00 (.02)	.00 (.02)	.00 (.02)	.00 (.01)	.00 (.01)
φ_{LA}	.03 (.14)	.06 (.64)	.02 (.61)	.03 (.62)	.01 (.60)	.01 (.57)
β	.24 (.27)	.26 (.26)	.22 (.26)	.23 (.25)	.23 (.22)	.23 (.24)
No DIF: AUD	.01 (.02)	.02 (.02)	.01 (.01)	.01 (.02)	.01 (.01)	.01 (.01)
φ_{LA}	.02 (.15)	.02 (.17)	.01 (.14)	.01 (.62)	.02 (.56)	.01 (.58)
β	.03 (.06)	.03 (.29)	.02 (.25)	.02 (.25)	.02 (.24)	.02 (.23)
Skew-2 g_R, Skew-1 g_F						
Unequal-a: AUD	.01 (.02)	.01 (.02)	.01 (.02)	.01 (.02)	.00 (.01)	.00 (.02)
β	.13 (.29)	.16 (.30)	.11 (.27)	.13 (.27)	.09 (.25)	.11 (.24)
Equal-a: AUD	.01 (.02)	.01 (.02)	.00 (.02)	.01 (.02)	.00 (.01)	.00 (.01)
φ_{LA}	.04 (.15)	.06 (.59)	.01 (.62)	.03 (.62)	.01 (.60)	.01 (.58)
β	.27 (.30)	.27 (.28)	.24 (.26)	.26 (.26)	.23 (.24)	.24 (.27)
No DIF: AUD	.02 (.02)	.02 (.02)	.01 (.01)	.01 (.02)	.01 (.01)	.01 (.01)
φ_{LA}	.01 (.14)	.02 (.16)	.01 (.15)	.02 (.62)	.01 (.58)	.01 (.61)
β	.05 (.08)	.06 (.31)	.03 (.28)	.04 (.27)	.02 (.23)	.02 (.25)

Note: k = number of items; g_R and g_F = latent distributions for reference and focal groups; AUD = average unsigned difference; φ_{LA} = Liu-Agresti effect size; β = poly-SIBTEST effect size; RMSE = root mean square error; DIF = differential item functioning; D-F = differentially functioning.

was elevated for items with equal-a and unequal-a DIF, but not DIF-free items, which is consistent with the inflated Type I error observed for poly-SIBTEST.

Discussion

IRT-LR-DIF was compared to poly-SIBTEST and the GMH and Mantel tests when the latent distribution was nonnormal for both groups but presumed normal for IRT-LR-DIF. Because they lack an explicit assumption about the normality of the latent variables, the nonparametric procedures and their effect sizes were expected to be more accurate than IRT-LR-DIF and the AUD. Contrary to expectation, results showed that although latent nonnormality detrimentally affected all three DIF testing procedures, IRT-LR-DIF was the most robust. Type I error for conditions with latent nonnormality was closer to the nominal level for IRT-LR-DIF than for the nonparametric procedures.

Of the methods compared, poly-SIBTEST displayed the largest Type I error. Poly-SIBTEST, like SIBTEST, was developed assuming the latent distribution is equal for the two groups

Table 5. Absolute Bias (and RMSE) in Effect Sizes; $N_R = 1{,}200$, $N_F = 800$; Empirically Observed Pattern of DIF in b_{ij}

	$k = 12$		$k = 24$		$k = 48$	
Proportion of D-F items	1/6	1/3	1/6	1/3	1/6	1/3
Normal g_R and g_F						
Unequal-a: AUD	.00 (.01)	.00 (.02)	.00 (.01)	.00 (.01)	.00 (.01)	.00 (.01)
β	.29 (.42)	.32 (.45)	.32 (.45)	.32 (.44)	.32 (.44)	.32 (.44)
Equal-a: AUD	.00 (.01)	.00 (.01)	.00 (.01)	.00 (.01)	.00 (.01)	.00 (.01)
β	.22 (.35)	.25 (.45)	.20 (.44)	.23 (.45)	.23 (.45)	.22 (.43)
No DIF: AUD	.01 (.02)	.01 (.01)	.01 (.01)	.01 (.01)	.01 (.01)	.01 (.01)
β	.01 (.05)	.02 (.37)	.00 (.34)	.00 (.47)	.00 (.46)	.00 (.46)
Skew-1 g_R, Skew-2 g_F						
Unequal-a: AUD	.01 (.02)	.01 (.02)	.00 (.02)	.01 (.02)	.00 (.01)	.00 (.01)
β	.29 (.44)	.28 (.44)	.31 (.44)	.31 (.45)	.34 (.46)	.33 (.48)
Equal-a: AUD	.00 (.01)	.01 (.02)	.00 (.01)	.00 (.02)	.00 (.01)	.00 (.01)
β	.20 (.33)	.19 (.43)	.20 (.46)	.21 (.43)	.22 (.48)	.22 (.44)
No DIF: AUD	.01 (.02)	.02 (.01)	.01 (.01)	.01 (.02)	.01 (.01)	.01 (.01)
β	.03 (.06)	.03 (.33)	.02 (.34)	.02 (.43)	.02 (.47)	.02 (.45)
Skew-2 g_R, Skew-1 g_F						
Unequal-a: AUD	.01 (.02)	.01 (.02)	.01 (.02)	.01 (.02)	.00 (.01)	.01 (.02)
β	.30 (.44)	.26 (.43)	.30 (.43)	.30 (.45)	.34 (.47)	.32 (.45)
Equal-a: AUD	.01 (.02)	.01 (.02)	.01 (.02)	.01 (.02)	.00 (.01)	.00 (.01)
β	.18 (.33)	.17 (.42)	.19 (.44)	.21 (.43)	.22 (.47)	.21 (.45)
No DIF: AUD	.02 (.02)	.02 (.02)	.01 (.01)	.02 (.02)	.01 (.01)	.01 (.02)
β	.05 (.08)	.06 (.33)	.03 (.32)	.04 (.43)	.02 (.45)	.02 (.45)

Note: k = number of items; g_R and g_F = latent distributions for reference and focal groups; AUD = average unsigned difference; β = poly-SIBTEST effect size; RMSE = root mean square error; DIF = differential item functioning; D-F = differentially functioning.

(Shealy & Stout, 1993, pp. 168-169). When the assumption is false, Type I error is inflated, which is why the regression correction (described in the introduction of this paper) was developed to reduce the error. However, the regression correction assumes approximate linearity, which requires some minimum number of anchors (maybe around 20 for binary items; Shealy & Stout, 1993, p. 170). In the present study, the latent mean was 0.5 lower for the F versus R group for all conditions, but in the nonnormal conditions, the R and F groups differed with respect to not only mean but also distributional shape. Thus, the pattern of Type I error inflation for poly-SIBTEST observed in the present study was likely due to the lack of (a) distributional equality in nonnormal conditions and (b) approximate linearity for the regression correction with smaller numbers of anchors.

There was also Type I error inflation for the GMH and the Mantel tests. Both of these are generalizations of the MH test for binary items, which has shown Type I error inflation when the latent distribution differs between groups. Holland and Thayer (1988) showed that including the single studied item in the matching criterion for the MH reduces the inflation *when the underlying model is Rasch*. Because the GMH and Mantel are extensions of the binary MH test, a similar property may hold for them. The Mantel test has been shown to conform to a Rasch model also: Master's (1982) partial credit model (Camilli & Congdon, 1999). Consistently (as mentioned in the introduction), the Mantel produces inflated Type I error when items truly vary in discrimination ability. In the present study, items varied in discrimination ability, which may

be the reason Type I error was inflated for the GMH and Mantel tests. Because the items in this simulation varied in discrimination, the data advantaged IRT-LR-DIF, but the data were generated to be realistic. IRT-LR-DIF may not have performed so well if the data had not been generated from exactly the item response model used to fit the data.

These simulations have revealed that although poly-SIBTEST and the GMH and Mantel tests are usually thought of as nonparametric and requiring few assumptions, they appear to be making some less obvious assumptions that are problematic with certain types of realistic data. All three methods seem to assume equal latent distributions between groups, and when that is violated, to require either equal discrimination ability over items (GMH and Mantel tests) or an adequate number of anchors for approximate linearity for the regression correction (poly-SIBTEST). Remember also that all three are designed to detect only DIF with respect to thresholds (not discrimination) parameters.

The effect of latent nonnormality on poly-SIBTEST and the GMH and Mantel tests may also have been affected by the summed-score distributions. The latent and summed score distributions are not, in general, the same shape, but the latent distribution does influence the score distribution. In this study, score distributions for conditions with nonnormal latent distributions tended to have more categories with small frequencies (e.g., 2 or 3) compared to conditions with normal latent distributions. Computations within small-N score groups are less reliable, and more of them existed in conditions with latent nonnormality. It is unclear how best to handle small-N score groups; exclusion seems reasonable, but when many score groups are excluded, bias could result from the loss of information. In this simulation, poly-SIBTEST was programmed to exclude results from the calculation of $\hat{\beta}$ for score categories with two or fewer observations, and all data were used for the GMH and Mantel tests. In future research, it would be useful to explore alternative computational strategies for the nonparametric measures when there are numerous small-N score groups.

In conditions with controlled Type I error, statistical power was always quite high for IRT-LR-DIF and the GMH test, and usually quite high for poly-SIBTEST and the Mantel test in the special case of uniform DIF with the constant DIF pattern. For IRT-LR-DIF and the GMH test, power was even higher for the empirically observed versus constant DIF pattern, whereas for poly-SIBTEST and the Mantel, power usually decreased for the empirically observed versus constant DIF pattern. This is consistent with the findings of Zwick et al. (1993) for the Mantel versus GMH and is probably because poly-SIBTEST and the Mantel test evaluate a net signed DIF effect whereas IRT-LR-DIF and the GMH test evaluate a global, unsigned DIF effect.

Latent nonnormality had no discernible influence on the accuracy and variability of the effect sizes. AUD performed well in all conditions and was the most widely applicable effect size. The φ_{LA} was not evaluated for items with unequal a_is because there is not, at present, a definition of the true φ_{LA} for that case. It would be useful to focus future work on establishing a definition of the true φ_{LA} for items with unequal a_i, and then evaluating bias and RMSE. Because poly-SIBTEST is designed to identify DIF in the b_{ij}s when the a_is are equal, it makes sense that bias in β would be higher for items with unequal versus equal a_i; however, this was only true for the empirical versus constant pattern of DIF in the b_{ij}s.

Data simulated for the present study are realistic for many types of psychological and educational scales, but data from a single simulation study cannot emulate all possible tests encountered by practitioners. For example, some exams are longer, have fewer items with DIF, or are less discriminating than those studied here. It would be useful to repeat the simulations with other types of data. The nonparametric approaches may show greater robustness with longer scales having few DIF items so that the matching criterion is composed of many items and is more reliable. When the latent distributions differ between groups, poly-SIBTEST should

perform better when there are more anchor items, and the GMH and Mantel test should perform better when items do not vary in discrimination ability.

It would also be useful in future research to examine alternative shapes for the latent distributions. Only two particular (positively skewed and leptokurtic) curves were used as latent distributions in this study. As explained in the methods section, these shapes are realistic for some types of variables, but others surely occur. It would also be interesting to separate the effects of skewness from kurtosis in future research.

An additional question for future research is whether IRT-LR-DIF would remain preferable in the presence of both latent nonnormality and IRF misspecification. Bolt's (2002) study focused on perfect latent normality combined with egregious IRF misspecification, and the present study focused on perfect IRF specification with fairly egregious latent nonnormality. In practice, it is not uncommon to observe a mild violation of latent normality and a mild misfit of the graded model—it would be interesting to see how the methods compare in this case.

One alternative way to carry out IRT-LR-DIF is to estimate the latent densities instead of presuming they are normal. Such procedures are becoming available, and extant work shows that Type I error can be improved when the densities are estimated as empirical histograms (Woods, 2008b, in press) or spline-based densities (Woods, 2010). The PARSCALE program (Muraki & Bock, 2003) implements an IRT-based DIF-testing procedure in which the latent distributions for both groups as estimated as empirical histograms (Bock, Muraki, & Pfeiffenberger, 1988; Muraki, 2003, pp. 560-561). However, the procedure is not IRT-LR-DIF because there are no anchors; the latent scale is linked across groups by a constraint which implies that the overall difficulty of the test is the same for both groups. That constraint may be overly restrictive for some tests.

In the present study, anchors were presumed to be accurately specified. In other words, items presumed DIF-free were actually DIF-free. With real data, it can be difficult to select a perfectly DIF-free set of anchors. Previous simulation results suggest that a contaminated anchor set produces Type I error inflation (Stark et al., 2006; Wang & Yeh, 2003; Woods, 2009). Thus, in practice, Type I error may be even higher than values reported here. Various methods for empirically selecting anchors have been suggested and appear to help reduce Type I error inflation (Holland & Thayer, 1988; Kim & Cohen, 1995; Miller & Oshima, 1992; Navas-Ara & Gómez-Benito, 2002; Park & Lautenschlager, 1990; Woods, 2009).

Some recommendations for practitioners may be made on the basis of these results. First, it is noteworthy that if the latent densities are approximately normal, all the procedures are likely to produce accurate and reasonably powerful results if only 1/6 of the items function differently. However, in the case of latent nonnormality or a greater proportion of D-F items, it is risky to rely on poly-SIBTEST, the GMH, or the Mantel test when the number of anchors is less than about 12. By contrast, IRT-LR-DIF can perform well with as few as 3 anchors. If latent nonnormality is suspected, it appears to be better to use IRT-LR-DIF than any of these nonparametric procedures. Even better would be to use a variation of IRT-LR-DIF in which the latent densities are estimated from the data (e.g., Woods, 2008b). Because increasing the number of anchors had a mitigating effect on Type I error inflation for all methods, all items that can justifiably be presumed DIF-free should be included in the matching criterion. The most consistently accurate, widely applicable effect size was the AUD, an effect size associated with IRT-LR-DIF. Use of the AUD is recommended.

Author's Note

Carol M. Woods is now at the University of Kansas.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article:

The author was supported by NSF Grant SES-0818722.

Notes

1. Type I error can also be inflated for the three-parameter logistic model with small (e.g., 250 per group) sample sizes (Cohen, Kim, & Wollack, 1996).
2. A condition with ½ of the items differentially functioning was also included (fully crossed with the other factors), but results from it are not reported because an anonymous reviewer thought this condition was rare in practice, and the manuscript needed to be shortened. These results are available on request from the author.

References

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*, 277-300.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.

Bolt, D. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*, 113-141.

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.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 24*, 323-341.

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

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.

Cohen, A. S., Kim, S., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*, 15-26.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Hill, C. D. (2004). *Precision of parameter estimates for the graded item response model* (Unpublished master's thesis). University of North Carolina at Chapel Hill.

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

Kim, S., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education, 8*, 291-312.

Kim, S., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345-355.

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

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-143.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*, 91-100.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items and magnitude of bias on a two stage item bias estimation method. *Applied Psychological Measurement, 16*, 381-388.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E. (2003). 7.3 Models in PARSCALE. In M. du Toit (Ed.), *IRT from SSI* (pp. 544-566). Lincolnwood, IL: Scientific Software International.

Muraki, E., & Bock, R. D. (2003). PARSCALE-4: IRT item analysis and test scoring for rating-scale data [Computer software]. Chicago, IL: Scientific Software International.

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment, 18*, 9-15.

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.

Park, D., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14*, 163-173.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement, 29*, 150-151.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*, 187-210.

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

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.

Roussos, L., & Stout, W. (2005). Dimensionality-based DIF package: Poly-SIBTEST (version 1.7) [Computer software]. Urbana-Champaign, IL: William Stout Institute for Measurement.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.

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.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician, 40*, 106-108.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1291-1306.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*, 402-415.

Su, Y., & Wang, W. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18*, 313-350.

Sweeney, K. P. (1996). *A Monte-Carlo investigation of the likelihood-ratio procedure in the detection of differential item functioning* (Unpublished doctoral dissertation). Fordham University, New York.

Thissen, D. (2001). IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Documentation for computer program]. L.L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.

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

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123-135). Hillsdale, NJ: Erlbaum.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*, 197-219.

Wang, W., & Su, Y. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*, 450-480.

Wang, W., & Yeh, Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498.

Welch, C. J., & Hoover, H. D. (1993). Procedures for extending item bias techniques to polytomously scored items. *Applied Measurement in Education, 6*, 1-19.

Woods, C. M. (2006). Ramsay-curve item response theory to detect and correct for nonnormal latent variables. *Psychological Methods, 11*, 253-270.

Woods, C. M. (2008a). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement, 32*, 511-526.

Woods, C. M. (2008b). IRT-LR-DIF with estimation of the focal-group density as an empirical histogram. *Educational and Psychological Measurement, 68*, 571-586.

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

Woods, C. M. (2010). *Ramsay-curve differential item functioning*. Manuscript submitted for publication.

Woods, C. M. (in press). DIF testing with an empirical-histogram approximation of the latent density for each group. *Applied Measurement in Education*.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71*, 281-301.

Zwick, R., Donoghue, J., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233-251.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*, 321-344.