Likelihood-Ratio DIF Testing: Effects of Nonnormality

Carol M. Woods, Washington University in St. Louis

Differential item functioning (DIF) occurs when an item has different measurement properties for members of one group versus another. Likelihood-ratio (LR) tests for DIF based on item response theory (IRT) involve statistically comparing IRT models that vary with respect to their constraints. A simulation study evaluated how violation of the normality assumption about the random latent variable for one or both groups affected IRT-LR-DIF results. Item response data with or without DIF were generated from the two-parameter logistic model and fitted under the assumption that the latent distribution was normal for both groups. Although the IRT-LR-DIF method performed well when latent distributions were normal for both groups, results were distorted when the distribution was skewed for one or both groups. Specifically, Type I error was inflated, differences between reference- and focal-group item parameter estimates were inaccurate, and group differences in the mean and variance of the latent distribution were overestimated. *Index terms: differential item functioning, LR-DIF, IRT-LR-DIF, item response theory, item bias, measurement invariance*

Items with differential item functioning (DIF) have different measurement properties for one group of people than another, irrespective of group-mean differences on the variable under study. Detecting DIF is important because it can mislead researchers about group differences and invalidate procedures for making decisions about individuals (for insightful work on the relationship between DIF and validity, see Borsboom, 2006; Borsboom, Mellenbergh, & van Heerden, 2002). Numerous methods have been proposed for identifying DIF (Camilli & Shepard, 1994; Holland & Wainer, 1993; Millsap & Everson, 1993). The present research is concerned with the LR method based on item response theory (IRT), abbreviated IRT-LR-DIF (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993). When IRT-LR-DIF procedures are carried out using marginal maximum likelihood estimation (MML; Bock & Aitkin, 1981; Bock & Lieberman, 1970), in which the latent variable (θ) is treated as random, it is usually assumed that θ is normal for both groups. The purpose of the present research is to evaluate how MML IRT-LR-DIF results can be affected by a violation of this normality assumption for one or both groups.

DIF

In an IRT context, DIF means that an item has a different item response function (IRF) for one group of people than another. In other words, even when members of two different groups are matched on θ, their probability of giving the same response to a particular item is not the same. Groups are defined by, for example, sex, ethnicity, or experimental condition, with one assigned to

Applied Psychological Measurement, Vol. 32 No. 7, October 2008, 511–526
DOI: 10.1177/0146621607310402

be the *reference group* and the other assigned to be the *focal group*. The majority, or group with which a test was originally developed, may be the reference group, whereas the group for whom items are expected to show DIF is the focal group.

DIF testing can be carried out for items with binary, ordinal, or nominal response options, with the probability of the responses modeled by any of various IRFs. Here, the focus is on binary items, modeled with the popular two-parameter logistic (2PL; Birnbaum, 1968) IRF. The 2PL discrimination parameter (a) indicates how well an item distinguishes between people with higher versus lower levels of θ, and the 2PL threshold parameter (b) is the value of θ at which the discrimination occurs. The threshold is also the value of θ at which a respondent is equally likely to respond in one category versus the other.

A statistically significant test of DIF with respect to a indicates that an item is less discriminating for one group than the other. This has been referred to as *nonuniform DIF* (Camilli & Shepard, 1994, p. 59; Mellenbergh, 1989) because the nature of the DIF varies over the range of θ. DIF in a (DIF-a) indicates that the IRFs for the two groups cross at some value of θ. Therefore, the item is more easily endorsed by the reference group at certain ranges of θ, but more easily endorsed by the focal group at other ranges of θ. DIF-a can occur when an item measures a different latent variable for one group than the other; the meaning of the latent variables can be examined in research subsequent to DIF analyses (e.g., with structural equation modeling).

Interpretation of a test for DIF with respect to b is most straightforward when as are constrained equal between groups so that it may be assumed that the DIF (if present) is constant over the range of θ and that the same latent variable is being measured for both groups. A test for DIF in b, conditional on equal as, is a test of *uniform DIF* (Camilli & Shepard, 1994, p. 59). A statistically significant result indicates that the level of θ at which the response categories are equally likely to be endorsed differs for the two groups. The expression "DIF-b|equal-a," used in this article to represent uniform DIF, refers to the presence of DIF in the b parameter conditional on equal a parameters.

LR-DIF Testing Based on IRT (IRT-LR-DIF)

In IRT-LR-DIF, several two-group IRT models, varying in their constraints, are statistically compared. This is done separately for every studied item (i.e., item to be tested for DIF). Some items, called *anchors*, must be assumed DIF-free and are used to set a common scale for θ. The parameters of anchor items are constrained equal between groups whereas studied items are evaluated for DIF. Typically, the θ distribution, $g(\theta)$, is assumed normal for both groups with the mean and standard deviation (SD) fixed at 0 and 1, respectively, for the reference group and estimated for the focal group simultaneously with the item parameters.

For binary items fitted with the 2PL IRF, three DIF tests are of interest for each studied item. First is an omnibus test for DIF in a and b simultaneously, carried out for a single item at a time. This test is "omnibus" because it could be significant if there is DIF in a, DIF in b, or both. The null (H_o) and alternative (H_a) hypotheses are

$$\begin{aligned} &H_o: a_F = a_R \text{ and } b_F = b_R \text{ for item } i, \\ &H_a: \text{ not all item parameters for item } i \text{ are group equivalent,} \end{aligned} \tag{1}$$

where F represents the focal group and R represents the reference group. To carry out this test, a model with both parameters for the studied item constrained equal between groups is compared to a model with both parameters for the studied item permitted to vary between groups. In both models, both parameters for all anchor items are constrained equal between groups. The (approximately) χ^2-distributed test statistic is -2 times the difference between the optimized log likelihoods, with degrees of freedom

(*df*) equal to the difference in the number of free parameters. With two parameters per item, $df = 2$. Statistical significance indicates the presence of DIF.

Following a significant omnibus test, more specific tests of uniform and nonuniform DIF may be carried out. Both are χ^2-difference tests between nested models, with $df = 1$ (when the 2PL IRF is used). Hypotheses for an LR test of nonuniform DIF (i.e., DIF with respect to the *a* parameter) are

$$\begin{aligned} &H_o: a_F = a_R \text{ and } b_F \neq b_R \text{ for item } i, \\ &H_a: a_F \neq a_R \text{ and } b_F \neq b_R \text{ for item } i. \end{aligned} \tag{2}$$

The models being compared differ in whether the *a* parameter for the studied item is permitted to vary between groups. In both models, the *b* parameter is free to vary between groups for the studied item, and both *a* and *b* are constrained equal between groups for all anchor items. The *b* parameter is free to vary between groups because if the item has DIF in *a*, it may be measuring a different latent variable in one group versus another, in which case there would be no justification for assuming the *b*s are equal. Statistical significance indicates the presence of nonuniform DIF.

The null hypothesis for the test of uniform DIF is the same as that for the omnibus test, but H_a differs:

$$\begin{aligned} &H_o: a_F = a_R \text{ and } b_F = b_R \text{ for item } i, \\ &H_a: a_F = a_R \text{ and } b_F \neq b_R \text{ for item } i. \end{aligned} \tag{3}$$

For the test of uniform DIF, the models being compared differ in whether *b* for the studied item is permitted to vary between groups. In both models, *a* is constrained to be group equivalent for the studied item, and both *a* and *b* are constrained to be group equivalent for all anchor items. The test of uniform DIF is a test of DIF with respect to *b*, conditional on the absence of DIF in *a*. Therefore, it is reasonable only when the test for nonuniform DIF is nonsignificant. Statistical significance indicates the presence of uniform DIF.

Previous Research

In simulation research with DIF-free studied items, Type I error (i.e., the probability of rejecting a true null hypothesis) for the omnibus IRT-LR-DIF test has been well controlled in various situations, as long as most anchor items are actually DIF-free. Type I error rates have been close to the nominal level for 2PL, three-parameter logistic (3PL), and graded models (Ankenmann, Witt, & Dunbar, 1999; Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998; Sweeney, 1996; Wang & Yeh, 2003), and the empirical mean and *SD* of the LR statistic have been near what they should be for a χ^2-distributed statistic (Ankenmann et al., 1999). Also, the focal-group mean has been recovered well when it differed from the reference-group mean by 1 *SD* (Kim & Cohen, 1998). Wang and Yeh (2003) found that if at least 12% (for the 3PL model) or 20% (for the 2PL or graded models) of anchor items actually had DIF, Type I error was inflated to levels such as .08 or .12 ($\alpha = .05$), with inflation worsening as the percentage of anchor items with DIF increased.

In simulation research with DIF in the data and DIF-free anchor items, power to detect threshold DIF (difference between bs $= .25$) in five-category items fitted with the graded model was sometimes high (e.g., .80 or larger) but depended on the magnitude of the true *a* parameters (1.7 versus 1.0), the sample size, and whether the mean of $g(\theta)$ differed between groups (by 1 *SD*). Optimal conditions for maximizing power were larger *a*s, larger *N*s, and identical population means (Ankenmann et al., 1999). In a separate study, Wang and Yeh (2003) found that power to detect omnibus DIF that consistently favored one group (rather than sometimes favoring one

group and sometimes favoring the other group) was high with the 2PL, 3PL, and graded models. Reference and focal *b* parameters differed by .3 or .4, and *a* parameters differed by .3.

All simulation studies reviewed above were carried out with normal $g(\theta)$ for both groups. As argued elsewhere (Woods, 2006a; Woods & Thissen, 2006), this normality assumption may be unrealistic, especially for latent variables pertaining to personality or psychopathology. Simulations outside the context of DIF indicate that IRT item parameters can be biased when $g(\theta)$ is nonnormal but assumed normal when the model is fitted (Boulet, 1996; De Ayala, 1995; Kirisci & Hsu, 1995; Stone, 1992; van den Oord, 2005; Yamamoto & Muraki, 1991; Zwinderman & van den Wollenberg, 1990). It is difficult to imagine that IRT-LR-DIF results would not also be affected. However, exactly how they can be affected is unclear.

The Present Study

This article describes a simulation study carried out to evaluate the effect of skewness of $g(\theta)$ on the (a) Type I error rate, (b) power, (c) degree to which the mean and *SD* of the test statistic match those for the corresponding χ^2-distributed statistic, (d) degree to which accurate amounts of DIF are detected, and (e) accuracy of estimated group differences in the mean and *SD* of θ. IRT-LR-DIF assuming normal $g(\theta)$ will be evaluated with $g(\theta)$ either skewed for both groups, skewed for one group, or normal for both groups (for comparison). Whereas previous simulations have focused on either the omnibus DIF test or a test of DIF in a single parameter (e.g., only thresholds; Ankenmann et al., 1999), the present study includes an evaluation of omnibus DIF tests as well as the follow-up tests described above.

Method

Design and Summary

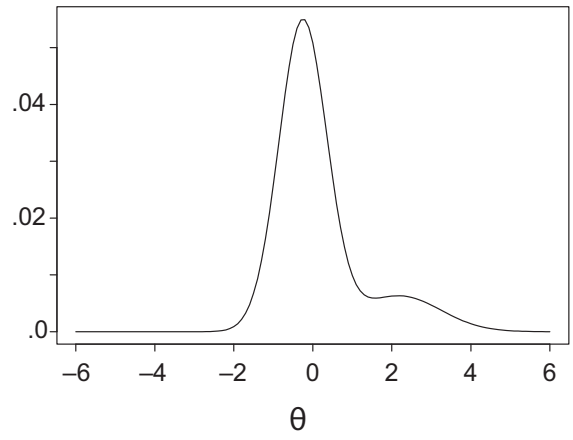
A $2 \times 2 \times 4$ factorial design produced 16 simulation conditions. Manipulated variables were (a) the presence of DIF in the data (present/absent), (b) whether there was a mean difference on θ between groups (yes/no), and (c) the distribution of θ (four patterns, explained below). For each condition, 1,000 data sets were generated. Each data set consisted of item response patterns for 24 binary items generated for 2,000 simulees (1,500 in the reference group and 500 in the focal group). In the 8 conditions for which DIF was present, 9 items were DIF-free anchors, 5 items had DIF in the *a* (but not *b*) parameter, and 10 items had DIF in the *b* (but not *a*) parameter. In the 8 conditions for which there was no DIF, parameters for all 24 items were identical for the reference and focal groups. Further details and justifications for these choices are provided below.

True Distributions of $g(\theta)$

Four distributional patterns for θ were evaluated. First, $g(\theta)$ was normal for both groups. This provided a baseline condition indicating how well the method performs when the data and the model match. Second, $g(\theta)$ was normal for the reference group but skewed for the focal group. Third, $g(\theta)$ was normal for the focal group but skewed for the reference group. Fourth, $g(\theta)$ was skewed for both groups.

The skewed curve is plotted in Figure 1. It is a mixture of two normals ($\mu_1 = -0.253$, $\mu_2 = 2.192$, $\sigma_1 = 0.609$, $\sigma_2 = 1.045$, $mp_1 = .897$, $mp_2 = .103$; mp = mixing proportion) with skewness and kurtosis coefficients equaling 1.57 and 3.52, respectively. (Three has been subtracted from the kurtosis

Figure 1
Skewed Curve Used as a True Latent Distribution in the Simulations



coefficient for interpretability; this coefficient is 0 for a perfectly normal distribution.) These levels of skewness and kurtosis are similar to those estimated previously with real data using a method that estimates the distribution nonparametrically using basis B-splines (see Woods, 2006a; Woods & Thissen, 2006). A latent distribution estimated based on binary items related to panic disorder had skewness and kurtosis coefficients equaling 1.04 and 3.53 (Woods & Thissen, 2006).[1]

Decisions Based on Published Applications of IRT-LR-DIF

Published applications of IRT-LR-DIF were used to inform choices required for the simulation. A sample of studies reporting sufficient detail (e.g., differences between reference- and focal-group item parameters) were used (Balsis, Gleason, Woods, & Oltmanns, 2007; Bielinski & Davison, 1998; Chan, Orlando, Ghosh-Dastidar, Duan, & Sherbourne, 2004; Donovan & Drasgow, 1999; Ellis, Becker, & Kimmel, 1993; Ellis & Kimmel, 1992; Mackinnon et al., 1995; Morales, Reise, & Hays, 2000; Oishi, 2006; Orlando & Marshall, 2002; Reise, Widaman, & Pugh, 1993; Rodebaugh, Woods, Heimberg, Liebowitz, & Schneier, 2006; Smith & Reise, 1998; Stark, Chernyshenko, Chang, Lee, & Drasgow, 2001; Steinberg, 2001; Williams, Turkheimer, Schmidt, & Oltmanns, 2005).

Sample sizes. The mean reference-group *N* reported in this literature (excluding three unusually large ones) was 1,081. The reference group was commonly larger than the focal group; considering studies for which this was true, the mean ratio of reference to focal *N* was approximately 3 (Balsis et al., 2007; Donovan & Drasgow, 1999; Ellis & Kimmel, 1992; Orlando & Marshall, 2002; Rodebaugh et al., 2006; Williams et al., 2005). In the present simulations, the reference-group *N* (1,500) was three times the focal-group *N* (500).

Scale length. The number of items in a unidimensional set analyzed with IRT-LR-DIF was typically between 7 and 38, with a mean over studies (excluding a few outliers) of 20. The scale length for the present simulations was 24.

Mean differences on θ. The group difference between means on the latent variable, $|\bar{\theta}_R - \bar{\theta}_F|$, estimated as part of the DIF analyses, was reported in six studies (Morales et al., 2000; Oishi,

2006; Orlando & Marshall, 2002; Reise et al., 1993; Rodebaugh et al., 2006; Smith & Reise, 1998). The average difference was 0.6. Group differences in the variance of θ were rarely reported.

In the present simulations, a group-mean difference on θ was a manipulated variable. For the eight simulation conditions with no mean difference, the true θ distributions were constructed to have $\mu = 0$ and $\sigma = 1$ for both groups. For the eight conditions with a mean difference, $g(\theta_R)$ was standardized ($\mu = 0$ and $\sigma = 1$), but the mean of $g(\theta_F)$ was -0.6. The variance of θ was always 1.

Nature and amount of DIF. In several applications, only omnibus DIF was tested; thus, results of more specific tests (for uniform and nonuniform DIF) were not available. When specific tests were done, statistically significant DIF was usually due to differences in the *b* parameter (Balsis et al., 2007; Chan et al., 2004; Morales et al., 2000; Orlando & Marshall, 2002; Rodebaugh et al., 2006; Smith & Reise, 1998; Steinberg, 2001). Significant DIF in the *a* parameter was sometimes observed, for about 15% to 20% of items (Morales et al., 2000; Rodebaugh et al., 2006). The amount of DIF always varied over items within a study. Typical threshold differences, $|b_F - b_R|$, were between .3 and .7. Typical slope differences, $|a_F - a_R|$, were also in this range.

In the present study, data with DIF were generated so that 5 of 24 items had nonuniform DIF and 10 of 24 items had uniform DIF. The amount of DIF in a given item was randomly selected to be .3, .4, .5, .6, or .7. All simulated DIF favors the reference group: Items with nonuniform DIF are less discriminating for the focal group (i.e., $a_F < a_R$), and items with uniform DIF are more difficult for the focal group to answer correctly or endorse (i.e., $b_F > b_R$). The 9 DIF-free items are used as anchors to set a common scale for θ between the two groups and are never tested for DIF.

Anchor items. Post hoc equating (Stocking & Lord, 1983) has been used in lieu of anchor items in some applications of IRT-LR-DIF (Donovan & Drasgow, 1999; Ellis & Kimmel, 1992; Ellis et al., 1993; Mackinnon et al., 1995; Stark et al., 2001). However, in other research, anchors have been used and explicitly mentioned (Balsis et al., 2007; Oishi, 2006; Orlando & Marshall, 2002; Reise et al., 1993; Rodebaugh et al., 2006) or not explicitly mentioned but, in the present author's judgment, used (Smith & Reise, 1998; Steinberg, 2001). In practice, anchor items may be assumed DIF-free based on previous research, expert opinion, or a purification analysis prior to the DIF analysis (e.g., see Kim & Cohen, 1995). Alternatively, IRT-LR-DIF analyses may be carried out treating all other items as anchors (i.e., using all items except the one studied item to equate the scale).

The present simulations use a particular set of correctly specified anchor items. In previous simulations with normal $g(\theta)$, Wang and Yeh (2003) showed that the Type I error rate in IRT-LR-DIF testing is inflated when the anchor set includes at least 12% to 20% incorrectly specified anchors (i.e., items with DIF). Thus, in practice, it is important to identify anchor items as correctly as possible, but this issue is not the focus of the present study. In applications of IRT-LR-DIF, the number of anchor items has been 25% to 41% of the total (Balsis et al., 2007; Orlando & Marshall, 2002; Reise et al., 1993). Here, the number of DIF-free anchors was nine (37.5% of the total).

Data Generation

For all tests, 24 binary item responses were generated from the 2PL model. For the reference group, *a*s were randomly drawn from a normal distribution ($\mu = 1.7$, $\sigma = 0.3$), chosen based on an empirical examination of discrimination parameters estimated from an assortment of psychological scales (Hill, 2004), with truncation to avoid unrealistic extreme values. The distribution of true

a parameters was truncated on the upper end at 4 and at the lower end at 0.5 (conditions without DIF) or 1.2 (conditions with DIF). The maximum amount of DIF was .7, so truncation at 1.2 ensured that the focal-group *a* was never smaller than 0.5. Threshold parameters were randomly drawn from a normal distribution ($\mu = 0$, $\sigma = 1$) and truncated at ± 2 to avoid items with all responses in a single category. Focal-group parameters were defined in relation to reference-group parameters as described in the previous section.

Model Fitting and DIF Testing

Parameter estimation and IRT-LR-DIF testing were carried out using C++ source code from the IRT-LR-DIF program (Version 2.0b; Thissen, 2001), with a simulator (in C++) added for this study. No modifications to the estimation procedures in Thissen's code were made. Bock and Aitkin's (1981) scheme for MML implemented with an expectation-maximization (EM) algorithm was used to fit the 2PL model to the item response vectors. The θ distribution was assumed standard normal for the reference group and normal for the focal group, with mean and *SD* estimated simultaneously with the item parameters. The latent variable was represented with rectangular quadrature, ranging from -6 to 6 in increments of 0.1 (121 points). The maximum number of EM cycles was 1,000 for fittings with all item parameters constrained equal in both groups, and 500 for all other fittings. A fitting was declared converged when the parameter that was changing the most between EM cycles changed less than .0001.

Nine DIF-free items were used as anchors (i.e., their parameters were constrained equal for the two groups) in all fittings. The remaining 15 items were tested for DIF, one at a time. For each studied item, the omnibus DIF test was performed first. If this test was significant ($\chi^2 > 5.99$, $df = 2$, $\alpha = .05$),[2] the test for nonuniform DIF was carried out. If this test was nonsignificant ($\chi^2 < 3.84$, $df = 1$, $\alpha = .05$), the test for uniform DIF was carried out.

Outcome Measures

For conditions without DIF, the proportion of all tests with a significant DIF test (averaging over items and replications) was computed as a measure of Type I error. The empirical mean and *SD* of the LR statistic for each test was compared to the values they should be for a χ^2-distributed statistic: mean = *df*, *SD* = sqrt(2*df*).

For conditions with DIF, the proportion of replications with a significant DIF test was computed as a measure of Type I error for items without DIF and of power for items with DIF (Type I error and power will be computed separately for each of the three DIF tests). In addition to evaluating whether DIF is statistically significant, it is important to evaluate the degree to which an item shows DIF. A simple measure of the amount of DIF is the difference between reference and focal parameters, averaged over items. For items with DIF, the degree to which an accurate amount and type of DIF was observed was evaluated by comparing the difference between the reference and focal-group parameters (estimated from the model permitting them to vary between groups) to the true difference. To control outliers for this analysis, any *a* estimates greater than 4 were recoded to 4, and any *b* estimates outside the quadrature range (more extreme than ± 6) were recoded to ± 6.

The mean and *SD* of the focal $g(\theta)$ were approximated simultaneously with the item parameters. Because the mean is fixed at 0 for the reference group, the focal-group mean gives the group difference in means. For all conditions, the mean and *SD* of θ for the focal group (estimated from the model permitting item parameters to vary between groups) was compared to the true population parameters: $\mu = 0$ or -0.6 (depending on condition), and $\sigma = 1$.

Table 1
Distributional Characteristics of True θ and True Item Parameters: Differential Item Functioning (DIF)–Free Conditions

	Reference Group				Focal Group			
Condition	$\overline{\beta}_1$	$\overline{\beta}_2$	$\overline{a}$	$\overline{b}$	$\overline{\beta}_1$	$\overline{\beta}_2$	$\overline{a}$	$\overline{b}$
				Normal θ_R, normal θ_F				
$\overline{\theta}_R = \overline{\theta}_F$	0.00	0.00	1.70	0.01	0.00	0.00	1.70	0.01
$\overline{\theta}_R \neq \overline{\theta}_F$	0.00	0.00	1.70	0.00	0.00	0.01	1.70	0.00
				Skewed θ_R, skewed θ_F				
$\overline{\theta}_R = \overline{\theta}_F$	1.56	3.51	1.70	0.00	1.56	3.54	1.70	0.00
$\overline{\theta}_R \neq \overline{\theta}_F$	1.56	3.51	1.70	0.01	1.56	3.52	1.70	0.01
				Normal θ_R, skewed θ_F				
$\overline{\theta}_R = \overline{\theta}_F$	0.00	0.00	1.70	0.00	1.56	3.52	1.70	0.00
$\overline{\theta}_R \neq \overline{\theta}_F$	0.00	0.00	1.70	0.00	1.57	3.57	1.70	0.00
				Skewed θ_R, normal θ_F				
$\overline{\theta}_R = \overline{\theta}_F$	1.56	3.52	1.70	–0.01	0.00	0.00	1.70	–0.01
$\overline{\theta}_R \neq \overline{\theta}_F$	1.56	3.51	1.70	0.00	0.00	–0.01	1.70	0.00

Note. $\overline{\beta}_1$ = mean skewness coefficient for true θ; $\overline{\beta}_2$ = mean kurtosis coefficient for true θ; $\overline{a}$ = mean of true discrimination parameters for items without *a* DIF; $\overline{b}$ = mean of true threshold parameters for items without *b* DIF; θ_R = reference-group latent variable; θ_F = focal-group latent variable; $\overline{\theta}_R = \overline{\theta}_F$ = the mean of θ is fixed to 0 for both groups; $\overline{\theta}_R \neq \overline{\theta}_F$ = the mean of θ is 0 for the reference group and −0.6 for the focal group.

Results

Characteristics of the Simulated Data

Tables 1 and 2 present (average) skewness and kurtosis coefficients for the distribution of true θ values. Values are near those of the distribution from which the values were drawn. Also listed are the mean and *SD* of the true item parameters for each simulation condition (averaged over items and replications). For conditions with DIF (Table 2), descriptives for *a* parameters are given separately for the 19 items without DIF in *a* ($\bar{a}$) and the 5 items with DIF in *a* ($\bar{a}_{DIF}$). Analogously, descriptives for *b* parameters are given separately for items with ($\bar{b}_{DIF}$, 10 items) and without ($\bar{b}$, 14 items) DIF in *b*.

Convergence

Convergence was achieved in all replications for all 16 conditions.

Recoding of Extreme Slope Estimates

In each of the 16 simulation conditions, 15,000 *a*s were estimated separately for each group (15,000 = 15 studied items × 1,000 data sets). Only a small percentage of *a*s were greater than 4 (thus, they were recoded to 4). All conditions required recoding of some focal-group *a*s except the two with skewed θ_R, normal θ_F, and no group-mean difference. Only conditions with skewed θ_R

Table 2
Distributional Characteristics of True θ and True Item Parameters: Conditions With Differential Item Functioning (DIF)

	Reference Group				Focal Group					
Condition	$\overline{\beta}_1$	$\overline{\beta}_2$	$\overline{a}$	$\overline{b}$	$\overline{\beta}_1$	$\overline{\beta}_2$	$\overline{a}$	$\overline{a}_{DIF}$	$\overline{b}$	$\overline{b}_{DIF}$
					Normal θ_R, normal θ_F					
$\overline{\theta}_R = \overline{\theta}_F$	0.00	0.00	1.73	−0.01	0.00	−0.01	1.73	1.23	−0.01	0.49
$\overline{\theta}_R \neq \overline{\theta}_F$	0.00	0.00	1.73	0.00	0.01	−0.01	1.73	1.23	0.01	0.49
					Skewed θ_R, skewed θ_F					
$\overline{\theta}_R = \overline{\theta}_F$	1.56	3.52	1.73	0.01	1.56	3.54	1.73	1.23	0.00	0.52
$\overline{\theta}_R \neq \overline{\theta}_F$	1.56	3.50	1.73	0.01	1.56	3.50	1.73	1.24	0.02	0.51
					Normal θ_R, skewed θ_F					
$\overline{\theta}_R = \overline{\theta}_F$	0.00	0.00	1.74	0.00	1.55	3.51	1.74	1.24	0.01	0.48
$\overline{\theta}_R \neq \overline{\theta}_F$	0.00	0.00	1.73	0.01	1.56	3.50	1.73	1.23	0.00	0.51
					Skewed θ_R, normal θ_F					
$\overline{\theta}_R = \overline{\theta}_F$	1.56	3.52	1.73	0.00	0.00	0.00	1.73	1.23	−0.01	0.51
$\overline{\theta}_R \neq \overline{\theta}_F$	1.56	3.52	1.73	−0.01	0.00	0.00	1.73	1.23	−0.01	0.50

Note. $\overline{\beta}_1$ = mean skewness coefficient for true θ; $\overline{\beta}_2$ = mean kurtosis coefficient for true θ; $\overline{a}$ = mean of true discrimination parameters for items without *a* DIF; $\overline{a}_{DIF}$ = mean of true discrimination parameters for items with *a* DIF; $\overline{b}$ = mean of true threshold parameters for items without *b* DIF; $\overline{b}_{DIF}$ = mean of true threshold parameters for items with *b* DIF; θ_R = reference-group latent variable; θ_F = focal-group latent variable; $\overline{\theta}_R = \overline{\theta}_F$ = the mean of θ is fixed to 0 for both groups; $\overline{\theta}_R \neq \overline{\theta}_F$ = the mean of θ is 0 for the reference group and −0.6 for the focal group.

needed recoding of reference-group *a*s. Typically, more *a*s needed recoding in conditions with versus without DIF and with versus without a group-mean difference on θ.

When θ was normal for both groups, between 0.01% and 0.15% of *a*s were recoded for the focal group; none were recoded for the reference group. When θ was skewed for both groups, between 0.33% and 0.94% of focal-group *a*s were recoded, and between 0.11% and 0.13% of reference-group *a*s were recoded. When θ_R was normal and θ_F was skewed, between 0.55% and 1.83% of focal-group *a*s (and 0% of reference-group *a*s) were recoded. When θ_R was skewed and θ_F was normal, between 0.10% and 0.13% of reference-group *a*s were recoded, and results varied for the focal group. When there was no group-mean difference on θ, 0% of focal-group *a*s were recoded, but with a mean difference, 0.01% (no DIF) or 0.10% (DIF) of *a*s were recoded.

None of the estimated *b*s were outside the range ± 6; thus, none were recoded.

Results for Conditions Without DIF

The first row of Table 3 gives the Type I error rate for the omnibus test, and the second row gives the empirical mean and *SD* for the corresponding LR test statistic. When θ_R and θ_F were both normal, Type I error was at the nominal level, and the empirical mean and *SD* for the omnibus test statistic were near the values for the corresponding χ^2 distribution ($M = 2$, $SD = 2$). This was also true when θ_R and θ_F were both skewed and there was no mean difference on θ. When there was a mean difference, the Type I error rate and the mean and *SD* of the test statistic were a little

Table 3
Simulation Results for Differential Item Functioning (DIF)–Free Conditions (1,000 Replications)

	$\overline{\theta}_R = \overline{\theta}_F = 0$				$\overline{\theta}_R = 0; \overline{\theta}_F = -0.6$			
	$\theta_R \sim N, \theta_F \sim N$	$\theta_R \sim S, \theta_F \sim S$	$\theta_R \sim N, \theta_F \sim S$	$\theta_R \sim S, \theta_F \sim N$	$\theta_R \sim N, \theta_F \sim N$	$\theta_R \sim S, \theta_F \sim S$	$\theta_R \sim N, \theta_F \sim S$	$\theta_R \sim S, \theta_F \sim N$
Sig. omnibus tests	.05	.05	.14	.14	.05	.08	.18	.13
M (*SD*) of the omnibus test statistic	1.98 (2.01)	1.92 (1.94)	2.99 (3.14)	3.10 (3.29)	1.97 (1.96)	2.42 (2.40)	3.50 (3.61)	3.00 (3.24)
Sig. DIF-*a* tests	.48	.45	.73	.70	.47	.45	.77	.73
M (*SD*) of the DIF-*a* test statistic	3.94 (3.09)	3.62 (3.02)	6.20 (3.91)	5.96 (4.00)	3.87 (2.94)	3.71 (3.12)	6.58 (3.96)	6.40 (4.18)
Sig. DIF-*b*lequal-*a* tests	.88	.94	.89	.88	.89	.92	.87	.84
M (*SD*) of DIF-*b*lequal-*a* test statistic	6.29 (2.23)	6.51 (2.01)	6.36 (2.41)	6.59 (2.94)	6.23 (2.05)	6.86 (2.50)	6.26 (2.58)	6.38 (3.13)
$\overline{\theta}_F$ (*SD*), n.s. omnibus	0.00 (.06)	0.00 (.06)	−0.07 (.05)	0.07 (.07)	−0.60 (.06)	−0.76 (.07)	−0.69 (.07)	−0.65 (.08)
SD of θ_F (*SD*), n.s. omnibus	1.00 (.05)	1.00 (.06)	0.86 (.07)	1.17 (.09)	1.00 (.06)	1.19 (.08)	0.98 (.09)	1.21 (.08)

Note. θ_R = reference-group latent variable; θ_F = focal-group latent variable; $\overline{\theta}_R$ = mean of θ_R; $\overline{\theta}_F$ = mean of θ_F; $\sim$N = normal; $\sim$S = skewed. The nominal alpha level for all tests was .05.

elevated. When θ was skewed for only one of the groups, Type I error and the mean and *SD* of the test statistic were quite inflated. Results were most biased when θ_R was normal, θ_F was skewed, and there was a group-mean difference on θ.

Tests of nonuniform (DIF-*a*) and uniform (DIF-*b*|equal-*a*) DIF were carried out only when the omnibus test was significant. In DIF-free conditions, a significant omnibus test is a Type I error. Thus, if a Type I error was made for the omnibus test, the Type I error rate (and test statistic) were substantially elevated for the more specific tests. The Type I error rate for DIF-*a* tests when either θ_R or θ_F was skewed was nearly double the rate observed when the distribution of θ_R and θ_F was the same. The Type I error for DIF-*b*|equal-*a* tests was high for all conditions but was highest when θ_F and θ_R were both skewed.

The last two rows of Table 3 list the estimated mean and *SD* of θ_F (with an *SD* for each, indicating variability over replications) when the omnibus test was correctly found to be nonsignificant. If both θ_F and θ_R were normal, the focal-group mean and *SD* were estimated well. This was also true if θ_F and θ_R were both skewed, but only when there was not a mean difference on θ. The focal-group mean and *SD* were estimated less accurately in the other five simulation conditions. The mean was usually underestimated. The *SD* was underestimated when θ_R was normal (with skewed θ_F) and overestimated when θ_R was skewed (with normal θ_F).

Results for Conditions With DIF

The first row of Table 4 shows that the proportion of significant omnibus tests (i.e., power) was fairly high for all conditions. However, the high power for many of these conditions is partially due to Type I error inflation. The second row of Table 4 lists the proportion of significant tests of nonuniform DIF (DIF-*a*) for items with DIF in *b*. Because these items did not have DIF with respect to the *a* parameter, this proportion indicates Type I error. Type I error was at the nominal rate for the normal–normal conditions, but elevated in the presence of any skewness. The third row of Table 4 gives the proportion of significant DIF-*a* tests for items with DIF in *a* (i.e., power), which are fairly high for all conditions.

Next, Table 4 lists the proportion of significant tests of uniform DIF (DIF-*b*|equal-*a*) for items with DIF in *a* (fourth row) or DIF in *b* (fifth row). This test was carried out only when the omnibus test was significant and the DIF-*a* test was nonsignificant; thus, it is not surprising that the test was almost always significant, irrespective of the latent distributions.

For items with DIF in *a*, the true *a* for the focal group was, on average, 0.5 smaller than that for the reference group, and the *b* parameters were identical for the two groups. As shown in Table 4 (sixth row), the average difference between reference- and focal-group *a*s was usually somewhat overestimated, but was quite underestimated when θ_R and θ_F were both skewed. The mean difference between *b*s was accurate for the normal–normal conditions but inaccurate in the presence of any skewness (Table 4, row 7).

For items with DIF in *b*, the true focal- and reference-group *a* parameters were identical, and the true focal-group *b* was, on average, about 0.5 larger than that for the reference group. As shown in Table 4 (eighth row), if θ_F was normal (with θ_R either normal or skewed), the difference between *a*s was estimated fairly well; otherwise, it was quite inaccurate. If both θ_F and θ_R were normal, the difference between *b*s was estimated fairly well, but accuracy was poor in the presence of any skewness (Table 4, row 9).

The last two rows of Table 4 list the estimated mean and *SD* of θ_F (with an *SD* for each, indicating variability over replications) when the omnibus test was correctly found to be significant. The same pattern of results observed for conditions without DIF was observed for conditions with DIF.

Table 4
Simulation Results for Conditions With Differential Item Functioning (DIF) Favoring the Reference Group (1,000 Replications)

	$\overline{\theta}_R = \overline{\theta}_F = 0$				$\overline{\theta}_R = 0; \overline{\theta}_F = -0.6$			
	$\theta_R \sim$ N, $\theta_F \sim$ N	$\theta_R \sim$ S, $\theta_F \sim$ S	$\theta_R \sim$ N, $\theta_F \sim$ S	$\theta_R \sim$ S, $\theta_F \sim$ N	$\theta_R \sim$ N, $\theta_F \sim$ N	$\theta_R \sim$ S, $\theta_F \sim$ S	$\theta_R \sim$ N, $\theta_F \sim$ S	$\theta_R \sim$ S, $\theta_F \sim$ N
Sig. omnibus tests	.85	.82	.85	.85	.84	.82	.84	.86
Sig. DIF-*a* tests for items with DIF in *b*	.05	.11	.22	.16	.05	.12	.25	.13
Sig. DIF-*a* tests for items with DIF in *a*	.82	.69	.72	.77	.81	.80	.74	.80
Sig. DIF-*b*|not *a* tests for items with DIF in *a*	.90	.95	.94	.96	.93	.95	.96	.93
Sig. DIF-*b*|not *a* tests for items with DIF in *b*	1	1	1	1	1	1	.99	1
$a_R - a_F$ (sig. omnibus) for items with DIF in *a*	.62	.27	.58	.58	.62	.37	.57	.61
$b_R - b_F$ (sig. omnibus) for items with DIF in *a*	.00	−.15	.11	−.13	−.01	−.11	.12	−.14
$a_R - a_F$ (sig. omnibus) for items with DIF in *b*	−.02	−.24	.23	−.02	−.03	−.18	−.26	.02
$b_R - b_F$ (sig. omnibus) for items with DIF in *b*	−.53	−.57	−.39	−.68	−.55	−.59	−.38	−.73
$\overline{\theta}_F$ (*SD*) (sig. omnibus)	0.00 (.06)	0.00 (.06)	0.07 (.05)	0.07 (.07)	−0.61 (.06)	−0.76 (.07)	−0.68 (.06)	−0.65 (.08)
SD of θ_F (*SD*) (sig. omnibus)	1.00 (.06)	1.00 (.06)	0.86 (.07)	1.16 (.09)	1.00 (.06)	1.20 (.08)	0.98 (.09)	1.21 (.09)

Note. *R* = reference group; *F* = focal group; ∼N = normal; ∼S = skewed. The nominal alpha level for all tests was .05.

Discussion

Key Findings

Results for the normal–normal simulation conditions support the validity of the IRT-LR-DIF method and are consistent with previous simulation studies using normal $g(\theta)$ and correctly specified anchor items (Ankenmann et al., 1999; Cohen et al., 1996; Kim & Cohen, 1998; Sweeney, 1996; Wang & Yeh, 2003). The IRT-LR-DIF program (Version 2.0b; Thissen, 2001) is easily obtained (from Thissen), but other programs such as MULTILOG (Thissen, 1991), BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), or Mplus (Muthén & Muthén, 2006) can be used for IRT-LR-DIF testing. The main advantage of the IRT-LR-DIF program is convenience: A single run of the program tests each individual item for DIF, and it is easy to either specify particular anchor items or use all other items as anchors. IRT-LR-DIF testing with other currently available programs requires many separate runs to test each item for DIF.

The present study contributes some specific findings about how IRT-LR-DIF results can be misleading when the normality assumption about θ is violated. Problems created by skewness included elevation in Type I error rates and misestimation of group differences in the item parameters. In other words, skewness can lead researchers to detect DIF when there is none present and either under- or overestimate the amount of DIF. This can needlessly slow down the test development or refinement process if many DIF-free items are rewritten because they are (falsely) believed to show DIF. Skewness also caused overestimation of group differences on the mean and variance of θ, which confuses efforts to understand substantive differences between reference and focal groups, separate from the issue of DIF.

One unexpected finding was that IRT-LR-DIF was relatively robust to misspecification of the latent distributional shape, provided that the shape, mean, and variance of θ were identical for the two groups. Perhaps when the populations are identical, bias in the MML results is constant across groups so that DIF tests are minimally affected. In practice, it is probably rare for the shape, mean, and variance of θ to be identical for the reference and focal groups. Thus, this special case of robustness is probably not very common with empirical data sets, and any skewness should be cause for concern.

Limitations and Future Research

Efforts were made to generate realistic data, but all simulations are limited to some degree so that the number of conditions is manageable. One potentially important variable that was not manipulated is the direction of the DIF. Wang and Yeh (2003) pointed out that it is unlikely in practice that all items favor the same group. In the present study, all DIF favored the reference group. It would be interesting to see if these results hold when DIF sometimes favors the reference group and other times favors the focal group.

Many questions remain that could be addressed in subsequent simulations of nonnormal θ in IRT-LR-DIF. Speculations above predict that any differences in the latent distributions between groups would produce problems such as those observed here. Thus, bias would be expected if, for example, $g(\theta)$ was nonnormal but not the same shape in both groups, or if the nonnormal shape and mean of $g(\theta)$ were the same in both groups but the variance differed between groups. The degree of bias in the IRT-LR-DIF results might increase as group differences in $g(\theta)$ increase.

This study could be repeated with different kinds of shapes for the true θ distributions. Previous research suggests that MML IRT (for a single group) is fairly robust to misspecification of $g(\theta)$ when the distribution is platykurtic rather than skewed (Woods & Thissen, 2006), but skewness has been consistently problematic. IRT-LR-DIF results may be more biased as the degree of

skewness increases. True shapes for unobservable latent distributions are largely speculative at present because methods for estimating them are relatively new or infrequently applied.

Perhaps more useful than additional simulations of this sort would be the development of methods and software for carrying out IRT-LR-DIF without the normality assumption. A few methods exist for estimating $g(\theta)$ simultaneously with the item parameters for a single sample. The most well known is the empirical histogram method (Bock & Aitkin, 1981; Mislevy, 1984) implemented in the BILOG-MG 3 program (Zimowski et al., 2003). Another (less flexible) approach is Thissen's (1991) Johnson-curve method implemented in the MULTLOG program (studied by van den Oord, 2005). A newer method called Ramsay-curve IRT estimates the latent density as a smooth, B-spline–based density (Woods, 2006a; Woods & Thissen, 2006) and is implemented in the RCLOG program (Version 1, Woods & Thissen, 2004; Version 2, Woods, 2006b). It would be useful to extend these methods for use in IRT-LR-DIF testing.

Notes

1. In the article by Woods and Thissen (2006), kurtosis coefficients were reported without subtracting 3; thus, the kurtosis coefficient for the normal distribution was 3. In the present article, 3 has been subtracted from the coefficient so that 0 is the kurtosis coefficient for the normal distribution. This explains the discrepancy between the kurtosis coefficient for the estimated θ distribution as reported here versus in the article by Woods and Thissen.
2. Thissen's IRT-LR-DIF software (Version 2.0b, Thissen, 2001) performs the test for nonuniform DIF if the omnibus test is larger than 3.84 (the $\alpha = .05$ critical value for the χ^2 distribution with 1 *df*), regardless of the actual *df* for the omnibus test. This is done because there is no possibility that any subsequent single-*df* tests will be significant if the omnibus test does not exceed the 1-*df* critical value, and the use of 3.84 reduces computational time (see Thissen, 2001, p. 9). In the present research, the IRT-LR-DIF code was unchanged, but the results were analyzed using the actual critical value for the omnibus test (5.99, $df = 2$, $\alpha = .05$).

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.

Balsis, S., Gleason, M. E., Woods, C. M., & Oltmanns, T. F. (2007). Age group bias in DSM-IV personality disorder criteria: An item response analysis. *Psychology and Aging*, *22*, 171-185.

Bielinski, J., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal*, *35*, 455-476.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison & Wesley.

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., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *35*, 179-197.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*, S176-S181.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement*, *26*, 433-450.

Boulet, J. R. (1996). The effect of non-normal ability distributions on IRT parameter estimation using full-information and limited-information

methods (item response theory, nonlinear factor analysis). *Dissertation Abstracts International*, *58*, 1256.

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

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) scale: An item response theory analysis. *Medical Care*, *42*, 281-289.

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.

De Ayala, R. J. (1995, April). *Item parameter recovery for the nominal response model.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Donovan, M. A., & Drasgow, F. (1999). Do men's and women's experiences of sexual harassment differ? An examination of the differential test functioning of the sexual experiences questionnaire. *Military Psychology*, *11*, 265-282.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology*, *24*, 133-148.

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

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., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence 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.

Kirisci, L., & Hsu, T.C. (1995, April). *The robustness of BILOG to violations of the assumptions of unidimensionality of test items and normality of ability distribution*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Mackinnon, A., Jorm, A. F., Christensen, H., Scott, L. R., Henderson, A. S., & Korten, A. E. (1995). A latent trait analysis of the Eysenck personality questionnaire in an elderly community sample. *Personality and Individual Differences*, *18*, 739-747.

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

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297-334.

Mislevy, R. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359-381.

Morales, L. S., Reise, S. P., & Hays, R. D. (2000). Evaluating the equivalence of health care ratings by Whites and Hispanics. *Medical Care*, *38*, 517-527.

Muthén, L. K., & Muthén, B. O. (2006). Mplus: Statistical analysis with latent variables (Version 4.1) [Computer software]. Los Angeles, CA: Authors.

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality*, *40*, 411-423.

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.

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.

Rodebaugh, T. L., Woods, C. M., Heimberg, R. G., Liebowitz, M. R., & Schneier, F. R. (2006). The factor structure and screening utility of the social interaction anxiety scale. *Psychological Assessment*, *18*, 231-237.

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

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

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology*, *81*, 332-342.

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

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameters logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*, 1-16.

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. (1991). MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory [Computer software and manual]. Chicago: Scientific Software International.

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 [Computer software documentation]. Chapel Hill: L. L. Thurstone Psychometric Laboratory, University of North Carolina.

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. 147-170). Hillsdale, NJ: Lawrence 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: Lawrence Erlbaum.

van den Oord, E. J. C. G. (2005). Estimating Johnson curve population distributions in MULTILOG. *Applied Psychological Measurement*, *29*, 45-64.

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.

Williams, M. T., Turkheimer, E., Schmidt, K. M., & Oltmanns, T. F. (2005). Ethnic identification biases responses to the Padua Inventory for Obsessive Compulsive Disorder. *Assessment*, *12*, 174-185.

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

Woods, C. M. (2006b). *RCLOG v.2: Software for item response theory parameter estimation with the latent population distribution represented using spline-based densities* (Tech. Rep.). St. Louis: Washington University.

Woods, C. M., & Thissen, D. (2004). *RCLOG v.1: Software for item response theory parameter estimation with the latent population distribution represented using spline-based densities* (Tech. Rep.). Chapel Hill: L. L. Thurstone Psychometric Laboratory, University of North Carolina.

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

Yamamoto, K., & Muraki, E. (1991, April). *Nonlinear transformation of IRT scale to account for the effect of non-normal ability distribution on the item parameter estimation*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). BILOG-MG 3 [Computer software]. Lincolnwood, IL: Scientific Software International.

Zwinderman, A. H., & van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement*, *14*, 73-81.

Acknowledgment

The author is grateful to Andrew Martin for comments on an early version of this article.

Author's Address

Address correspondence to Carol M. Woods, Psychology Department, Campus Box 1125, Washington University in St. Louis, St. Louis, MO 63130; e-mail: cwoods@artsci.wustl.edu.