Article

The Langer-Improved Wald Test for DIF Testing With Multiple Groups: Evaluation and Comparison to Two-Group IRT

Educational and Psychological Measurement
73(3) 532–547

DOI: 10.1177/0013164412464875
epm.sagepub.com

Carol M. Woods[1], Li Cai[2], and Mian Wang[1]

Abstract

Differential item functioning (DIF) occurs when the probability of responding in a particular category to an item differs for members of different groups who are matched on the construct being measured. The identification of DIF is important for valid measurement. This research evaluates an improved version of Lord's χ^2 Wald test for comparing item response model parameter estimates between two groups. The improved version uses better approaches for computation of the covariance matrix and equating the item parameters across groups. There are two equating algorithms implemented in IRTPro and flexMIRT software: Wald-1 (one-stage) and Wald-2 (two-stage), only one of which has been studied in simulations before. The present study evaluates for the first time the Wald-1 algorithm and Wald-1 and Wald-2 for three groups simultaneously. A comparison to two-group IRT-LR-DIF is included. Results indicate that Wald-1 performs very well and is recommended, whereas Type I error is extremely inflated for Wald-2. Performance of IRT-LR-DIF and Wald-1 was similar, even for three groups.

Keywords

differential item functioning, item response theory, Wald test, IRT-LR-DIF

[1]University of Kansas, Lawrence, KS, USA
[2]University of California, Los Angeles, Los Angeles, CA, USA

Corresponding Author:
Carol M. Woods, Department of Psychology, University of Kansas, 1415 Jayhawk Blvd., Room 426, Lawrence, KS 66045-7556, USA.
Email: cmw@ku.edu

Differential item functioning (DIF) occurs when the probability of responding in a particular category to an item differs for members of different groups who are matched on the construct being measured. There may be a true group-mean difference on the construct, but this is separate from whether the item functions differently between groups. Identification of DIF is vital for valid measurement of constructs in education, psychology, and other fields. Items that function differently for different groups may be eliminated or revised, or at least analyzed (and scored) as if a different item was administered to each group.

There are many methods available for testing DIF. The present study is concerned with two item response theory (IRT) approaches. One is the currently popular IRT-LR, also known as *IRT-LR-DIF* or *2-group IRT* (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993), in which nested two-group IRT models are compared with likelihood ratio tests. The other is an improved version of Lord's (1980, p. 212-224) Wald (1943) χ^2 test, a statistic comparing parameter estimates for an item between the reference and focal groups, divided by the standard error (*SE*) of their difference. For models with more than one item parameter, all parameters are compared simultaneously and the comparison is a matrix equation (given later in this article) with a covariance matrix instead of an *SE*. In Lord's (1980) original test, the item parameters are estimated separately in each group and the metric is subsequently equated using, for example, the approach of Stocking and Lord (1983).

Although IRT-LR and the Wald test are asymptotically equivalent (Thissen et al., 1993), the Wald test never performed very well in simulations, often showing severe Type I error inflation (Donoghue & Isham, 1998; Kim, Cohen, & Kim, 1994; Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987). Some authors suggested that inaccuracy in estimation of the covariance matrix was the problem (Donoghue & Isham, 1998; Kim et al., 1994; McLaughlin & Drasgow, 1987). Also, ad hoc equating does not always work as well as identification of the metric between groups while simultaneously estimating parameters. There were approximately 10 years when no methodological work was published on the Wald test.

The Wald test was recently improved (Cai, 2012; Cai, Thissen, & du Toit, 2011; Langer, 2008), so that the covariance matrix is estimated more accurately, and the latent scale is held constant over groups simultaneously with item parameter estimation rather than through ad hoc equating. With these improvements, the Wald test is expected to perform better. Indeed, in one extant simulation evaluation of the improved test, Type I error was well controlled (Langer, 2008). Langer (2008) evaluated one of two equating algorithms currently implemented in IRTPro (Cai et al., 2011) and flexMIRT (Cai, 2012) software for the improved Wald test. Both equating algorithms will be evaluated in the current simulations.

With more than two groups to compare, the currently implemented improved Wald test can compare all groups simultaneously using a contrast matrix. IRT-LR could theoretically compare multiple groups simultaneously but would require a linear model for the item parameters, which is not implemented in currently available software. Even if implemented, multiple-group IRT-LR would be much more time

consuming than the Wald approach because many more models must be fitted. The improved Wald test has not been evaluated in simulations for more than two groups. In the present study, three-group improved Wald testing is compared with pairwise comparisons with IRT-LR. Three-group Wald is expected to better control Type I error and provide more power than multiple pairwise IRT-LR.

The primary purpose of this research is to evaluate the improved Wald test and compare it to IRT-LR for ordinal responses. Conditions with two and three groups are examined, two different equating algorithms for the improved Wald test are compared, and the sample size and the percentage of DIF items are varied. The simulations are described following descriptions of IRT-LR and Wald testing.

IRT-LR DIF Testing

IRT-LR is a procedure developed in the 1980s wherein likelihood ratio (LR) tests from nested two-group unidimensional IRT models are used to test for DIF for one item at a time (Thissen et al., 1986, 1988, 1993). For scales that measure more than one (primary) construct, (essentially) unidimensional sets of items are individually tested for DIF. Almost any item response function (IRF) can be used, and the IRF may vary over items in the same analysis.

No explicit estimation of the latent construct, θ, is needed; θ is a random latent variable treated as missing using Bock and Aitkin's (1981) scheme for marginal maximum likelihood implemented with an expectation maximization algorithm (EM MML). The mean and variance of θ are fixed to 0 and 1 (respectively) for the reference group to identify the scale and estimated for the focal group as part of the DIF analysis. Anchor items, presumed to be group invariant, are needed to link the metric of θ for the two groups. Item parameters for all anchors are constrained equal between groups in all models. The studied items are tested individually for DIF.

For each studied item, an analysis begins with a general test that is designed to identify both uniform and nonuniform DIF (Camilli & Shepard, 1994; Mellenbergh, 1989). An analysis of a studied item fitted with (for example) Samejima's (1969, 1997) graded model for Likert-type response scales 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 the following:

H_0: $a_{i\mathrm{F}} = a_{i\mathrm{R}}$ and $b_{ij\mathrm{F}} = b_{ij\mathrm{R}}$ for all j
H_a: not all parameters for item i are group invariant

where ''F'' is for *focal* and ''R'' is for *reference*.

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 negative twice the difference between the optimized log likelihoods, which is approximately χ^2-distributed with degrees of freedom (*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, unequal b_{ij}s, or both.

IRT-LR can be accomplished by repeatedly running any software that fits item response models and permits parameters to be constrained equal between groups for some items. Thissen's (2001; Version 2.0b) free computer program, IRTLRDIF, provides a convenient implementation of IRT-LR by performing all of the model fitting required to test each studied item for DIF (both the general test and the follow-up tests) in a single run. IRTLRDIF compares two groups at a time, and if more than two groups are of interest, the program must be rerun such that the comparisons are totally separate, and overlapping. It is possible to extend IRT-LR for multiple groups (tested simultaneously), but would require many model fittings. There is currently no available software that implements simultaneous comparisons among more than two groups using IRT-LR.

When assumptions are met (e.g., latent variables are actually normally distributed), Type I error for the general IRT-LR 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 2-parameter logistic, 3-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, one item, or 10%, 16%, 20%, or 40% of the total number of items.

The Wald Test for DIF

For the 2-parameter logistic (2PL) model,

$$T_i(u_i = 1|\theta) = \frac{1}{1 + e^{[-a_i(\theta - b_i)]}},$$

with θ = latent variable, u_i = item response, a_i = discrimination, and b_i = difficulty, Lord's statistic is

$$\chi_i^2 = \mathbf{v}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{v}_i, \tag{1}$$

where $\mathbf{v}_i^T = [\hat{a}_{F_i} - \hat{a}_{R_i}, \hat{b}_{F_i} - \hat{b}_{R_i}]$, $\boldsymbol{\Sigma}_i$ = covariance matrix for differences between these item parameters, and *df* equal to the number of parameters compared (per item) between groups (e.g., *df* = 2 for the 2PL). This test can readily be used with different

unidimensional item response models instead; the contents of $\mathbf{v}_i$ and $\boldsymbol{\Sigma}_i$ are adjusted accordingly. Lord originally used joint maximum likelihood to estimate item parameters, but EM MML is used now.

The improved Wald test (Cai, 2012; Cai et al., 2011; Langer, 2008) is expected to improve on Lord's (1980) original because the covariance matrix is estimated using the supplemented expectation maximization (SEM) algorithm (Cai, 2008; Meng & Rubin, 1991), and the latent scale is held constant over groups simultaneously with item parameter estimation rather than through ad hoc equating. The SEM algorithm is a strategy for calculating the information matrix (used for item parameter *SE*s) when an EM algorithm is used for parameter estimation. Calculation of *SE*s is not straightforward with EM algorithms because the full parameter information matrix is not a by-product of the estimation as it is with non-EM maximum likelihood estimation.

Langer (2008) introduced a two-stage equating procedure referred to here as Wald-2, and Cai et al. (2011) introduced a one-stage equating procedure referred to here as Wald-1. Both Wald-1 and Wald-2 link the metric across groups simultaneously with item parameter estimation and DIF testing and should therefore improve on ad hoc linking. Both Wald-1 and Wald-2 are implemented in IRTPro and flexMIRT and use SEM estimation for the covariance matrix.

Wald-2

Wald-2 (Langer, 2008) does not require designated anchors. In the first stage, a model is fitted wherein the reference group mean and *SD* are fixed to 0 and 1 (respectively) to identify the scale, the mean and *SD* of the focal group are estimated, and all of the item parameters are constrained equal between groups. In the second stage, a model is fitted with the focal mean and *SD* fixed to the values obtained in the first stage. This links the metric between groups, and then all item parameters are free to vary between groups. As a result, the statistic in Equation (1) can be computed for each item as the test for DIF. In IRTPro or flexMIRT, this approach is invoked using an option called ''test all items, anchor all items.''

An advantage of Wald-2 is that all items are tested for DIF, and no anchors need to be specified. A disadvantage of Wald-2 is that without designated anchors, the focal group mean and *SD* are estimated from a misspecified model if there is DIF that does not cancel out across items in the first stage. DIF contamination of the anchor set is likely to produce Type I error inflation and other inaccuracies.

In Langer's (2008) simulations with 5, 20, or 40 five-category (graded model) ordinal responses, Type I error was well controlled, even a little underestimated (e.g., .02-.03 with $\alpha = .05$), but this was probably because power was quite low. Twenty percent of items on all simulated tests functioned differently between groups, and items were discrepant between groups by a difference of .1 or .2 in thresholds (b_{ij}s) and a multiple of 1.25 or 0.875 in discrimination (a_i). Reference group a_i was generated from $N(\mu = 1.7, \sigma^2 = 0.3^2)$ and b_{i1} from $N(\mu = -1.5, \sigma^2 = 0.5^2)$, with subsequent

b_{ij}s incremented by a value drawn from $N(\mu = 1, \sigma^2 = 0.2^2)$. Sample sizes were equal for the reference and focal groups, $N = 250$ or 1,000, and the focal group mean was 0 or −.6. As will be detailed in the present method section, data will be generated with larger differences between groups to create a more realistic evaluation of Type I error for Wald-2.

Wald-1

Cai et al. (2011) introduced Wald-1, which, like most DIF procedures, requires user-specified anchor items. Anchor items can be specified based on prior research or prior testing (e.g., using Wald-2 or another method). Purification or empirical anchor selection methods are commonly discussed in the DIF literature and lead to more accurate Type I error rates when there are studied items that function differently (Kim & Cohen, 1995; Wang, 2004; Woods, 2009).

In Wald-1, a single model is fitted wherein the reference group mean and *SD* are fixed to 0 and 1 (respectively) to identify the scale, and the mean and *SD* of the focal group are estimated simultaneously with estimation of the item parameters. Item parameters are either constrained equal between groups (anchor items) or free to vary between groups (studied items). A Wald statistic is obtained for every studied item. Wald-1 has not been previously studied in simulations. In IRTPro or flexMIRT, this approach is invoked using an option called, ''test candidate items, estimate group difference with anchor items.'' A candidate item is another name for a studied item.

Multiple-Group Wald

Kim, Cohen, and Park (1995) introduced a generalization of Lord's (1980) statistic to compare more than two groups simultaneously:

$$Q_i = (\mathbf{C}\mathbf{v}_i)^T (\mathbf{C}\boldsymbol{\Sigma}_i\mathbf{C})^{-1} (\mathbf{C}\mathbf{v}_i), \quad (2)$$

where $\mathbf{v}_i$ and $\boldsymbol{\Sigma}_i$ are similar to those in Equation (1) except that they hold item parameters and covariances (respectively) for *all* groups, and $\mathbf{C}$ is a matrix of contrast coefficients specified by the analyst that determine how the parameters are compared across groups. This approach used ad hoc linking and classic methods for estimating $\boldsymbol{\Sigma}_i$. Kim et al. (1995) provided an empirical example but not simulations, and the method was apparently never studied in simulations.

Multiple-group Wald testing can be improved also, by using SEM covariances and Wald-1 or Wald-2 linking algorithms. These are generalizations of the two-group methods described above. In multiple-group Wald-2, the means and *SD*s for *all* focal groups are estimated in Stage 1 with *all* item parameters group equivalent. In Stage 2, the means and *SD*s for *all* focal groups are fixed, with *all* item parameters free to vary across groups so that DIF tests are obtained (Equation 2). Langer (2008) provided an empirical example of three-group Wald-2 but not simulations.

Multiple-group Wald-1 has not been studied previously. In multiple-group Wald-1, a single model is fitted wherein the means and *SD*s for *all* focal groups are estimated while the item parameters for anchors are constrained equal across *all* groups, and the item parameters for studied items are free to vary across *all* groups. DIF tests are provided by the statistic in Equation (2).

Simulation Study

A simulation study was carried out to evaluate how the improved Wald tests (Wald-1 and Wald-2) perform with five-category ordinal data and two or three groups, under conditions varying by sample size and percentage of differentially functioning items. Two-group IRT-LR was included for comparison.

Method

Three characteristics were varied: the number of groups (two or three), the sample size (equal larger, equal smaller, unequal larger, unequal smaller), and the percentage of differentially functioning items on each test (25% or 50%). Therefore, there were $2 \times 4 \times 2 = 16$ simulation conditions. There were 500 replications in each condition. Two groups constitute the standard DIF comparison, and three groups provide a simple multiple-group case not previously evaluated. The selected percentages of DIF items are realistic for what is observed in applied research (Bolt, Hare, Vitale, & Newman, 2004; Chan, Orlando, Ghosh-Dastidar, Duan, & Sherbourne, 2004, Huang, Church, & Katigbak, 1997; Rodebaugh, Woods, Heimberg, Liebowitz, & Schneier, 2006; Steinberg, 2001).

Sample sizes were as follows. Equal larger = 1,000 for all groups, equal smaller = 500 for all groups, unequal larger = 1,500/500 (two groups) or 1,500/500/500 (three groups), unequal smaller = 750/250 (two groups) or 750/250/250 (three groups). The total sample sizes were chosen to be realistic for research using item response modeling procedures. Unequal group sizes extend the methods of Langer (2008) who used only equal group sizes. The proportions in the unequal group sizes used here are realistic for contexts in which it is reasonable to observe a 60% majority group with other groups at 20% frequency, for example, Caucasian, Asian, and Latino ethnic groups in parts of the United States.

Tests were generated with 24 five-category ordinal items, using Samejima's unidimensional graded model. Test length was not a variable of interest in this research and 24 is a moderate number of items that is realistically observed on unidimensional tests in a variety of fields. If an item was created with a 0 response frequency in one or more of the five intended response categories, the categories were collapsed and the item was analyzed as a three- or four- category item. The number of category collapses required for each replication was recorded for each condition.

Details for the true parameters were selected to be realistic for DIF research based on a review of applications of IRT-LR (Woods, 2009). Reference-group

discrimination parameters ($a_{i\mathrm{R}}$) were drawn from $N(\mu = 1.7, \sigma^2 = 0.6)$ with truncation on the upper end at 4.0, and on the lower end at 0.8. Truncation at 4.0 prevented $a_{i\mathrm{R}}$ from becoming unrealistically large, and the maximum amount of DIF in a_i was 0.7 so truncation at 0.8 ensured that $a_{i\mathrm{F}}$ was never less than 0.1. The first reference group threshold, $b_{i1\mathrm{R}}$, was drawn from $N(\mu = -0.4, \sigma^2 = 0.9)$ with truncation at -2.5 and 1.5. Subsequent thresholds were created by adding a randomly drawn value, $d_{ih\mathrm{R}}$, to the immediately previous threshold (h counts differences between consecutive $b_{ij\mathrm{R}}$s, where $j = 1, 2, 3, 4$). The difference between adjacent $b_{ij\mathrm{R}}$s was drawn from $N(\mu = 0.9, \sigma^2 = 0.4)$, with truncation at 0.1 and 1.5.

The number of correctly specified (i.e., DIF-free) anchor items was always 8. The rest of the test consisted of either (a) 6 items with DIF + 10 DIF-free studied items in conditions with 25% DIF or (b) 12 items with DIF + 4 DIF-free studied items in the conditions with 50% DIF. DIF-free studied items were used to evaluate Type I error. With Wald-2, the anchors were also tested for DIF; thus, they were also used to evaluate Type I error. For items with DIF, $a_{i\mathrm{R}} > a_{i\mathrm{F}}$ and $b_{ij\mathrm{R}} < b_{ij\mathrm{F}}$. The amount of DIF was varied randomly among the DIF items on a test, among three realistic values roughly considered smaller, medium, and larger amounts of DIF: $\delta = .3, .5$, or .7, determined separately for each a_i and b_{ij} (and separately for each focal group for three-group conditions) by a random draw (x) from a uniform (0, 1) distribution such that:

if $(x \leq .33)$ then $\delta = .3$
else if $((x > .33)$ and $(x \leq .66))$ then $\delta = .5$
else if $((x > .66)$ and $(x \leq 1.0))$ then $\delta = .7$

For the reference group, θ_{R} was drawn from $N(0, 1)$. For the (first) focal group, θ_{F1} was drawn from $N(-.6, 1)$. With three groups, the second focal group θ_{F2} was drawn from $N(-.8, 1)$. The focal mean of $-.6$ is realistic and was used by Langer (2008), and the third mean of $-.8$ was selected to be a little distinguishable from the other focal group mean.

With three groups, we specified two nonorthogonal contrasts indicating that each focal group was compared with the reference group: $\begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$. This matrix is set up so that each row is a contrast and each column is a group (column 1 is the reference group). Although orthogonal contrasts have desirable statistical properties, there is little value in testing comparisons that are not actually of substantive interest just because they are orthogonal. Here we used nonorthogonal contrasts corresponding to the group comparisons that are usually of interest for DIF.

All procedures were carried out using C++. IRT-LR was implemented using modified C++ source code from IRTLRDIF software (Thissen, 2001; Version 2.0b), which carries out item parameter estimation using Bock and Aitkin's (1981) EMMML estimation scheme. For the present study, modifications to the source code were made for three-group conditions so that two-group IRT-LR was carried out once between the first focal group and the reference group, and again between the

third group (a second focal group) and the reference group, for every replication. IRT-LR was always run with 8 correctly specified (i.e., DIF-free) anchors.

The latent variable was represented with rectangular quadrature, ranging from −5.5 to 5.5 in increments of .1 (111 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.

The improved Wald tests were carried out using the C++ numeric engine of flexMIRT (Cai, 2012). For every replication, testing was done once with Wald-1 and again with Wald-2. Models were estimated using the Bock–Aitkin EM algorithm and rectangular quadrature, ranging from −6 to 6 in increments of .25 (49 points), and the same convergence criterion as IRTLRDIF. None of the procedures are particularly sensitive to the number of rectangular quadrature points, so the discrepancy between the quadrature representations for the two procedures is not expected to influence the results. The covariance matrix of the estimated item parameters was computed using the SEM algorithm described by Cai (2008) with a convergence criterion of .001 for the forced-EM iterations.

Outcomes

Three results were evaluated: (a) statistical power, (b) Type I error rate, and (c) estimates of the focal group mean and *SD* ($\bar{\theta}_{F1}$ and SD_{F1} and also $\bar{\theta}_{F2}$ and SD_{F2} with three groups) averaged over studied items and replications. Statistical power was calculated for each simulation condition as the number of items with significant DIF tests divided by the number of items with DIF. Additionally, the percentage of false positives (the Type I error rate) was calculated as the number of items with a significant DIF test divided by the number of DIF-free studied items (multiplied by 100%). A binomial confidence interval (CI) was used to aid interpretation. With a true Type I error rate of .05 and 500 replications, the CI indicates that values between .03 and .07 are expected [CI computations: $.05 \pm 1.96(\sqrt{(.05(.95))/500}) = .05 \pm .02$]. A comparison between the average latent mean and *SD* and the known true values indicates the amount of bias.

Results

Response category collapsing was most frequent when the sample size was unequal smaller (between 99% and 94% of replications required at least one collapse), followed by equal smaller, unequal larger, and was most rare for equal larger (between 62% and 84% of replications required at least one collapse). More collapsing was required per replication for three versus two groups, which makes sense because the more data generated, the greater the likelihood of a 0 frequency.

Of primary interest in this study was the performance of the improved Wald tests. Results for the latent means and *SD*s, power, and Type I error are in Tables 1 to 3

Table 1. Estimated Latent Mean and (Standard Deviation).

	IRT-LR		Wald-1		Wald-2	
Sample Size	2 Groups	3 Groups	2 Groups	3 Groups	2 Groups	3 Groups
25% DIF						
Equal larger	−.60 (1.00)	Focal: −.60 (1.00) Third: −.80 (1.00)	−.60 (1.00)	Focal: −.60 (1.01) Third: −.80 (1.01)	−.67 (.94)	Focal: −.67 (.96) Third: −.86 (.95)
Equal smaller	−.61 (1.00)	Focal: −.60 (1.00) Third: −.81 (1.00)	−.61 (1.00)	Focal: −.60 (1.01) Third: −.81 (1.01)	−.67 (.94)	Focal: −.66 (.95) Third: −.86 (.95)
Unequal larger	−.60 (1.00)	Focal: −.60 (1.00) Third: −.80 (1.03)	−.60 (1.00)	Focal: −.60 (1.00) Third: −.80 (1.01)	−.66 (.94)	Focal: −.66 (.94) Third: −.86 (.94)
Unequal smaller	−.60 (1.00)	Focal: −.60 (1.00) Third: −.81 (1.00)	−.60 (1.01)	Focal: −.60 (1.00) Third: −.81 (1.01)	−.66 (.94)	Focal: −.66 (.93) Third: −.86 (.94)
50% DIF						
Equal larger	−.60 (1.05)	Focal: −.60 (1.00) Third: −.80 (1.00)	−.60 (1.01)	Focal: −.60 (1.01) Third: −.80 (1.00)	−.73 (.89)	Focal: −.74 (.90) Third: −.93 (.89)
Equal smaller	−.60 (1.07)	Focal: −.60 (1.00) Third: −.79 (1.00)	−.60 (1.01)	Focal: −.60 (1.01) Third: −.81 (1.01)	−.74 (.89)	Focal: −.74 (.90) Third: −.93 (.89)
Unequal larger	−.60 (1.00)	Focal: −.60 (1.00) Third: −.80 (1.00)	−.60 (1.00)	Focal: −.61 (1.00) Third: −.80 (1.01)	−.74 (.87)	Focal: −.74 (.87) Third: −.92 (.88)
Unequal smaller	−.60 (1.00)	Focal: −.60 (1.00) Third: −.80 (1.00)	−.60 (1.00)	Focal: −.61 (1.01) Third: −.80 (1.01)	−.74 (.87)	Focal: −.74 (.88) Third: −.92 (.87)

Note. DIF = differential item functioning; IRT-LR = two-group item response theory. The true mean and *SD* were: −.60 (1.00) for focal with two or three groups and −.80 (1.00) for the second focal (third) group. Sample sizes were equal larger = 1,000 for all groups, equal smaller = 500 for all groups, unequal larger = 1,500/500 (two groups) or 1,500/500/500 (three groups), unequal smaller = 750/250 (two groups) or 750/250/250 (three groups).

Table 2. Proportion of Hits (Power).

	IRT-LR		Wald-1		Wald-2	
Sample Size	2 Groups	3 Groups	2 Groups	3 Groups	2 Groups	3 Groups
25% DIF						
Equal larger	.98	Focal: .99 Third: .98	.98	Focal: .98 Third: .98	.97	Focal: .97 Third: .97
Equal smaller	.93	Focal: .93 Third: .92	.91	Focal: .91 Third: .90	.90	Focal: .88 Third: .88
Unequal larger	.96	Focal: .96 Third: .95	.97	Focal: .97 Third: .96	.95	Focal: .93 Third: .93
Unequal smaller	.88	Focal: .88 Third: .86	.90	Focal: .90 Third: .88	.85	Focal: .83 Third: .82
50% DIF						
Equal larger	.98	Focal: .98 Third: .98	.98	Focal: .97 Third: .97	.93	Focal: .93 Third: .93
Equal smaller	.93	Focal: .93 Third: .92	.92	Focal: .91 Third: .90	.83	Focal: .84 Third: .83
Unequal larger	.96	Focal: .96 Third: .94	.97	Focal: .97 Third: .95	.90	Focal: .90 Third: .88
Unequal smaller	.87	Focal: .88 Third: .86	.90	Focal: .90 Third: .88	.79	Focal: .78 Third: .77

Note. DIF = differential item functioning; IRT-LR = two-group item response theory. Sample sizes were equal larger = 1,000 for all groups, equal smaller = 500 for all groups, unequal larger = 1,500/500 (two groups) or 1,500/500/500 (three groups), unequal smaller = 750/250 (two groups) or 750/250/250 (three groups).

(respectively), in the middle column for Wald-1, and in the far-right column for Wald-2. Wald-1 performed superbly: The means and *SD*s were nearly identical to the generating values (Table 1), power to detect DIF was high (Table 2), and Type I error was near the nominal level (Table 3).

The story was quite different for Wald-2: Latent means and *SD*s were somewhat inaccurate (Table 1) and Type I error was inflated, egregiously so with 50% DIF items. The problem with Wald-2 is that differently functioning items are assumed to be DIF-free in the first stage. In the presence of DIF (that does not cancel out at the scale level), this model is misspecified, driving up the Type I error rate. The misspecification becomes increasingly extreme as the percentage of differentially functioning items increases. Because of the Type I error inflation, the power results for Wald-2 are not very useful (Table 2).

Results for IRT-LR are reported in the far left columns of Tables 1 to 3. Latent means and *SD*s were generally accurate; however, because of a few less accurate *SD* estimates, Wald-1 was preferable (Table 1). The Type I error rate was well controlled (Table 3). Power was similar to that for Wald-1; however, Wald-1 provided slightly greater power when the sample sizes were unequal (Table 2).

Table 3. Proportion of False Alarms (Type I Error).

	IRT-LR		Wald-1		Wald-2	
Sample Size	2 Groups	3 Groups	2 Groups	3 Groups	2 Groups	3 Groups
25% DIF						
Equal larger	.05	Focal: .05 Third: .05	.04	Focal: .04 Third: .04	.14	Focal: .11 Third: .12
Equal smaller	.05	Focal: .05 Third: .05	.04	Focal: .04 Third: .04	.07	Focal: .06 Third: .07
Unequal larger	.05	Focal: .04 Third: .05	.05	Focal: .04 Third: .05	.11	Focal: .09 Third: .10
Unequal smaller	.06	Focal: .04 Third: .05	.05	Focal: .04 Third: .05	.07	Focal: .05 Third: .07
50% DIF						
Equal larger	.05	Focal: .05 Third: .05	.04	Focal: .03 Third: .03	.56	Focal: .52 Third: .49
Equal smaller	.06	Focal: .04 Third: .05	.04	Focal: .03 Third: .04	.30	Focal: .26 Third: .25
Unequal larger	.05	Focal: .04 Third: .05	.05	Focal: .05 Third: .05	.47	Focal: .40 Third: .37
Unequal smaller	.05	Focal: .05 Third: .04	.04	Focal: .05 Third: .05	.22	Focal: .20 Third: .18

Note. DIF = differential item functioning; IRT-LR = two-group item response theory. Sample sizes were equal larger = 1,000 for all groups, equal smaller = 500 for all groups, unequal larger = 1,500/500 (two groups) or 1,500/500/500 (three groups), unequal smaller = 750/250 (two groups) or 750/250/250 (three groups).

Discussion

In pursuit of valid measurement for all test takers, it is important to test for, identify, and deal with any items that perform differently between groups. A differentially functioning item may be deleted, revised, or modeled (and scored) as if it was a different item for each group. The present study provided a first evaluation of the Wald-1 DIF-testing algorithm, and the three-group simultaneous test with Wald-1 and Wald-2 algorithms. Wald-1 and Wald-2 are implemented in IRTPro and flexMIRT software.

Wald-1 performed well. This method requires the specification of designated anchor items and was tested in the present simulation with correctly specified (DIF-free) anchors. In practice, anchors must be empirically selected, and the accuracy of the selection will influence the performance of the test. Anchors can be selected based on prior research or prior testing. If prior testing is necessary, the two-stage Wald testing method (Wald-2) may be useful for this purpose. A future simulation aimed at evaluating the performance of Wald-2 for anchor selection would be useful. Wald-1 is recommended for DIF testing, with thoughtful selection of anchors.

Wald-2 performed poorly; latent means and *SD*s were somewhat inaccurate and Type I error for DIF testing was inflated. Wald-2 seems appealing because the analyst need not designate anchors and DIF tests are obtained for all items. But the cost

of these conveniences is Type I error inflation, which will be minimal with smaller percentages, and extreme for higher percentages, of differentially functioning items. Type I error inflation suggests bias when minimal or no bias actually exists, which can lead analysts to mishandle (e.g., discard) fair items. Wald-2 is not recommended unless it is used to select anchors for Wald-1.

Notably, Langer's (2008) evaluation of Wald-2 did not indicate Type I error inflation. This seems to be because the DIF effect sizes were rather small in her study; true differences between thresholds were .1 or .2, whereas here we used .3, .5, or .7. The discrimination parameter differences are more difficult to compare between studies because Langer used a multiplicative constant and we used a simple difference. However, the expected value of the group discrimination difference in Langer's study was .21 or .43 ($1.7 - [1.7 \times 1.25]$ or $[1.7 \times 0.875]$, 1.7 was the mean of Langer's reference group a_i distribution) whereas we used .3, .5, or .7. Also, Langer simulated tests with 20% of items differentially functioning versus 25% or 50% in the present study. It is reasonable to expect Wald-2 to manifest inflated Type I error whenever it is used in the context of adequate power.

With three groups, Wald-1 was expected to show greater power to detect DIF than IRT-LR because it compares the groups simultaneously using a single statistic and model rather than a completely separate comparison for a third group as in IRT-LR. However, power was nearly the same for Wald-1 and IRT-LR. Perhaps the Wald method would show a power advantage if there were more than three groups, or if orthogonal contrasts were used. Langer (2008) discusses orthogonal contrast matrices for improved Wald DIF testing. The contrast matrix used in the present study reflected comparisons usually of interest in DIF, but it was not orthogonal; its use may have made the Wald and IRT-LR procedures more similar to one another than would be the case if an orthogonal contrast matrix were used. Future exploration of Wald-1 with different contrast matrices would inform this question.

The present study has focused on an omnibus test of DIF (null hypothesis rejected if any of the parameter estimates differ between groups for the item tested). However, if this test is significant, follow-up tests are often of interest to determine which item parameters actually differ between groups. For IRT-LR, such tests are obtained straightforwardly by fitting models with additional constraints, and are implemented in IRTLRDIF (Thissen, 2001). For follow-up tests with Wald testing, Langer (2008) described and studied conditional tests which are also implemented in IRTPro and flexMIRT. The conditional tests were not evaluated in the present study for the sake of simplicity and brevity, but it would be useful to examine their statistical properties in future research.

Other DIF testing methods that compare three or more groups simultaneously are multiple indicator multiple cause models (MIMIC; e.g., Muthén, 1985, 1989) and the generalized Mantel–Haenszel method (GMH; Somes, 1986; Zwick, Donoghue, & Grima, 1993). Wald testing is currently considered superior to both. The GMH approach conditions on summed scores, and only tests for uniform DIF. MIMIC models use latent variables, but implementation is straightforward only when

nonuniform DIF is ignored. In the future, MIMIC-interaction models (Woods & Grimm, 2011), which test for nonuniform DIF, will likely become more readily estimable, and then a comparison between them and Wald testing will be important.

Authors' Note

The first author is jointly appointed in Psychology and the Center for Research Methods and Data Analysis at the University of Kansas.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Li Cai benefits financially from the sale of IRTPro and flexMIRT software.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by the Institute of Education Sciences (R305B080016 and R305D100039) and the National Institute on Drug Abuse (R01DA026943 and R01DA030466).

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.

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.

Cai, L. (2008). SEM of another flavor: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*, 309-329.

Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software International.

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.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement,* 22, 33-51.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits. *Journal of Cross-Cultural Psychology, 28*, 192-218.

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.

Kim, S., Cohen, A., & Kim, H. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement, 18*, 217-228.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*, 261-276.

Langer, M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Unpublished doctoral dissertation). University of North Carolina, Chapel Hill.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*, 164-174.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11*, 161-173.

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

Meng, X., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association, 86*, 899-909.

Muthén, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics, 10*, 121-132.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557-585.

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.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*, 1-100.

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.

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *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. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81*, 332-342.

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

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 [Computer program]. Chapel Hill: 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. 67-111). Hillsdale, NJ: Erlbaum.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is larger. *Transactions of the American Mathematical Society, 54*, 426-482.

Wang, W. (2004). Effects of anchor item methods on detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education, 72*, 221-261.

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.

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

Woods, C. M., & Grimm, K. J. (2011) Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*, 339-361.

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