Applied Psychological Measurement
http://apm.sagepub.com/

Empirical Selection of Anchors for Tests of Differential Item Functioning

Carol M. Woods

Applied Psychological Measurement 2009 33: 42 originally published online 14 March 2008
DOI: 10.1177/0146621607314044

Empirical Selection of Anchors for Tests of Differential Item Functioning

Carol M. Woods, Washington University in St. Louis

Differential item functioning (DIF) occurs when items on a test or questionnaire have different measurement properties for one group of people versus another, irrespective of group-mean differences on the construct. Methods for testing DIF require matching members of different groups on an estimate of the construct. Preferably, the estimate is based on a subset of group-invariant items called designated anchors. In this research, a quick and easy strategy for empirically selecting designated anchors is proposed and evaluated in simulations. Although the proposed rank-based approach is applicable to any method for DIF testing, this article focuses on likelihood-ratio (LR) comparisons between nested two-group item response models. The rank-based strategy frequently identified a group-invariant designated anchor set that produced more accurate LR test results than those using all other items as anchors. Group-invariant anchors were more difficult to identify as the percentage of differentially functioning items increased. Advice for practitioners is offered. *Index terms: differential item functioning, item bias, measurement invariance, item response theory, likelihood ratio*

Differential item functioning (DIF) occurs when items on a test or questionnaire have different measurement properties for one group of people versus another, irrespective of group-mean differences on the construct, θ. Methods for DIF testing (Camilli & Shepard, 1994; Holland & Wainer, 1993; Millsap & Everson, 1993) require matching members of different groups on an estimate of θ. Sometimes all items, or all items less one, are used to estimate θ, but it is best if the estimate is based on only group-invariant items. Subsets of items that are presumed invariant and used to estimate or define the matching criterion are designated anchors.

Ideally, designated anchors are declared invariant based on "extensive data-analytic and expert review" (Thissen, Steinberg, & Wainer, 1993, p. 103). However, extensive prior research on anchors is rarely, if ever, carried out. If designated anchors are used, they are typically selected empirically in preliminary analyses of the same data that are used for the main DIF testing. Various strategies for empirically identifying anchors have been published, but many are complicated, untested, or both.

This article suggests a quick and easy strategy for empirically selecting anchors and evaluates it with simulations. The general strategy is applicable to any method for DIF testing but is presented and assessed using likelihood ratio tests (LRTs) that compare nested two-group item response theory (IRT) models (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993). This method is sometimes referred to as IRT-LR-DIF but is labeled IRT-LRT throughout this article to distinguish the methodology from Thissen's (2001) IRTLRDIF computer program.

Applied Psychological Measurement, Vol. 33 No. 1, January 2009, 42–57
DOI: 10.1177/0146621607314044

IRT-Based LRTs for DIF

The IRT-LRT approach involves comparing nested two-group item response models with varying constraints to evaluate whether the response function for a particular item differs for the reference and focal groups. Almost any item response function (IRF) can be used, and the IRF may vary over items in the same analysis. Samejima's (1997) graded model for items with ordinal responses is used here because it is popular in applications of IRT-LRT.

No explicit estimation of θ is needed; θ is a random latent variable treated as missing using Bock and Aitkin's (1981) scheme for maximum marginal likelihood. 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. Anchors link the metric of θ for the two groups; item parameters for all anchors are constrained equal across groups in all models. The nonanchors, called *studied items*, are tested individually for DIF.

For each studied item, an analysis begins with a general test that is sensitive to both uniform and nonuniform DIF (Camilli & Shepard, 1994, p. 59; Mellenbergh, 1989). The general test for item i fitted with the graded model is a test for DIF in the discrimination parameter, a_i; the threshold parameters, b_{ij}s (j indexes thresholds); or both. The null (H_o) and alternative (H_a) hypotheses are

$$\mathrm{H}_o : a_{iF} = a_{iR} \text{ and } b_{ijF} = b_{ijR} \text{ for all } j,$$
$$\mathrm{H}_a : \text{not all parameters for item } i \text{ are group invariant},$$

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 −2 times the difference between the optimized log likelihoods, which is approximately χ^2 distributed with degrees of freedom 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 uniform or nonuniform. Because power and Type I error rates for the follow-up tests are highly dependent on those for the general test, only the general test is evaluated in the present study.

All Others as Anchors

Sometimes, all other items are used as anchors for each studied item. If there is no DIF in any item, IRT-LRT performs well with all others as anchors (Cohen, Kim, & Wollack, 1996; S. Kim & Cohen, 1998). However, if some items function differently, the anchor set is contaminated, which causes at least three problems: distributional misspecification, inaccuracies in parameter estimates, and inflated false discovery rates. The LR statistic may fail to follow a χ^2 distribution. The LR statistic is χ^2 distributed when the larger of the models being compared holds (Haberman, 1977; Maydeu-Olivares & Cai, 2006). However, with all others as anchors, the larger model fits less well as the number of differentially functioning (DF) items, or the magnitude of the DIF, increases.

With DIF in the data and all others as anchors, Wang (2004) observed inaccuracies in estimates of the item parameters and the group-mean difference on θ as well as overestimation of the amount of DIF in the data. These problems may account for the more frequently reported problem of inflated Type I error or false discovery rate (Finch, 2005; Meade & Lautenschlager, 2004; Stark, Chernyshenko, & Drasgow, 2006; Wang, 2004; Wang & Yeh, 2003). The inflation is greater as true differences between reference and focal parameters increase or as the number of DF items increases (Stark et al., 2006; Wang, 2004; Wang & Yeh, 2003).

For analyses aimed at detecting DF items, the issues described above are clearly problematic. However, all others as anchors can be used only to identify designated anchors. In analyses aimed at identifying invariant items, an elevated false discovery rate is not necessarily a problem because researchers need not detect *all* invariant items to be anchors (and other results are not used). Nevertheless, an inflated false discovery rate could be problematic if no items appear to be suitable anchors.

Suggestions for Empirically Selecting Anchors

Lord (1980) described the general idea of iteratively purifying the matching criterion for DIF testing. Others have recommended, and sometimes tested, specific purification procedures for Mantel-Haenszel (MH), logistic regression, or post hoc–equated IRT methods (Candell & Drasgow, 1988; Holland & Thayer, 1988; Kok, Mellenbergh, & van der Flier, 1985; Miller & Oshima, 1992; Navas-Ara & Gómez-Benito, 2002; Park & Lautenschlager, 1990). Simulations suggest that purification tends to reduce false positives (e.g., Miller & Oshima, 1992; Navas-Ara & Gómez-Benito, 2002).

For IRT-LRT, Thissen et al. (1988) originally recommended choosing anchors based on preliminary tests using the MH procedure. IRT-LRT was computationally time consuming and logistically difficult at the time. However, with the creation of IRTLRDIF (Thissen, 2001) software and faster computers, alternative approaches seem preferable because the MH method uses observed summed scores for matching rather than a latent variable model and is insensitive to nonuniform DIF.

Three iterative purification procedures that have been suggested for IRT-LRT involve selecting anchors from preliminary IRT-LRT testing with all others as anchors (D. M. Bolt, Hare, Vitale, & Newman, 2004; Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; S. Kim & Cohen, 1995). Two other tactics begin with tests of each item using a single anchor, with every item taking a turn as the anchor (Rensvold & Cheung, 2001; Wang, 2004). Thus, if the total number of items (k) on a scale is 20, every item is tested for DIF 19 times, leading to $k(k-1) = 380$ tests. It is unclear whether the benefits of these iterative procedures outweigh the additional time and effort they require.

All of the purification approaches referenced above rely on the binary outcomes of hypothesis tests, known to produce more than the nominal number of significant results for a given α level when the anchor set is contaminated. The false discovery rate is inflated further when the same items are tested repeatedly and may be so large that no items appear to be invariant. Finally, iterative procedures might produce anchors with differential functioning that averages out to be indistinguishable from the mean difference, rather than invariant anchors (Thissen et al., 1993).

Recently, Stark et al. (2006) suggested selecting a single anchor based on a simple, noniterative rule. Following a test of all items with all others as anchors, one item is selected that has “the highest [factor] loading” among “(presumably) unbiased” items (p. 1304). Incorporating the factor loading (which is analogous to item discrimination) is reasonable because the anchor defines θ for the DIF analysis, so the anchor should be highly related to θ. However, there is no known relationship between the magnitude of a_i and the amount of DIF. Also, it may be difficult to find an item with both a large a_i and a nonsignificant DIF test because IRT-LRT tests have greater power when a_i is larger (Ankenmann, Witt, & Dunbar, 1999). Therefore, although Stark et al.’s two criteria make sense, they are unlikely to be attainable with many empirical data sets.

A Quick and Easy Rank-Based Strategy for Empirically Selecting Anchors

If the DIF status of all items is initially unknown, using all others as anchors is perhaps the only reasonable way to commence IRT-LRT testing. This article proposes a quick and easy strategy for using these preliminary results to select designated anchors. It is similar to the two-stage approaches suggested previously for other types of DIF methods (e.g., Holland & Thayer, 1988; Miller & Oshima, 1992) but differs because anchors are chosen based on the relative magnitude of the test statistics in Stage 1, not binary-choice hypothesis tests. The LR statistic is used but not compared to any distribution, so it is irrelevant whether it is χ^2 distributed. The strategy stems from the idea that the magnitude of the LR statistic reflects the degree to which an item functions differently between two groups, with larger LR values indicating greater DIF.

The proposed rank-based strategy is to (a) test all items for DIF with IRT-LRT using all others as anchors; (b) compute the ratio of the LR statistic to the number of free parameters, f, for each item; (c) rank order the items based on the LR/f ratio; and (d) designate g items with the g smallest LR/f ratios to be anchors (g may be 1). One goal of the simulation study is to determine the optimal choice for g. The LR/f ratio, rather than the LR statistic alone, is used to permit inclusion of items with different numbers of response options, or items fitted with different response models, in the same analysis. Of course, if f is constant over items, only the LR statistics are needed.

The optimal number of anchors is unclear. Every additional anchor represents an opportunity for contamination; thus, a single anchor (selected sensibly) might be best. In systematic evaluations of IRT-LRT with a single invariant anchor, group differences in item parameters and the mean difference between groups were well recovered, and Type I error was well controlled (Stark et al., 2006; Wang, 2004; Wang & Yeh, 2003).

On the other hand, anchors define θ, so validity is likely better with more anchors, especially if the items discriminate highly over a wide range of θ. Furthermore, with study characteristics held constant at realistic values, increasing the number of invariant anchors increases power (Wang, 2004; Wang & Yeh, 2003). Perhaps power increases with more anchors because fewer item parameters have to be estimated. Diminished power for single anchors has also been attributed to increased sampling variability in the item parameters (Wang, 2004).

A simulation study was carried out to evaluate how frequently the rank-based strategy produced a set of group-invariant anchors and to compare IRT-LRT results obtained using all others as anchors to those using varying numbers of empirically selected designated anchors. A key goal is to determine the optimal number of anchors (g) for various conditions.

Simulation Method

A C++ program was written to generate 100 data sets for each of 21 independent conditions: 18 with DIF and 3 without DIF. Simulation details were guided by 18 applications of IRT-LRT that are flagged with an asterisk in the reference list. Studied variables were the number of items ($k = 10, 20$, or 40), percentage of DF items (0%, 20%, 50%, or 80%), and type of DIF (uniform or nonuniform). The sample size was 1,500 for the reference group and 500 for the focal group. The distribution of θ was $N(0, 1)$ for the reference group and $N(-0.4, 1)$ for the focal group. Responses to five-category items were generated from Samejima's (1997) graded model.

Item Parameter Distributions and Amount of DIF

Reference-group item parameters were randomly drawn from certain distributions: $N(\mu = 1.7, \sigma = 0.6)$ for a_{iR} with truncation on the upper end at 4.0 and on the lower end at 0.5 (conditions without nonuniform DIF) or 1.2 (conditions with nonuniform DIF), and $N(\mu = -0.4, \sigma = 0.9)$ for b_{i1R} (with truncation at -2.5 and 1.5). The distance between consecutive b_{ijR}s, d_{imR} (where m counts differences between b_{ijR}s), was drawn from $N(\mu = 0.9, \sigma = 0.4)$ with truncation at 0.1 and 1.7.

Focal-group parameters were defined in relation to reference-group parameters. Items with uniform DIF had group-variant b_{ij} and items with nonuniform DIF had group-variant a_i and b_{ij}. For items with DIF in a_i, one of five equally likely values (.3, .4, .5, .6, or .7) was subtracted from a_{iR} to create a_{iF}. The pattern of DIF in b_{ij}s was selected based on the magnitude of the randomly drawn b_{i1R} from among nine patterns observed in applications of IRT-LRT (Orlando & Marshall, 2002; Reise, Widaman, & Pugh, 1993). The patterns and criteria used for selecting them are listed in Table 1. This procedure can produce true b_{ijF}s that are unordered (i.e., sets for which this is untrue: $b_{i1F} < b_{i2F} < b_{i3F} < b_{i4F}$). When an unordered set of b_{ijF}s was produced for item i, the C++ program rejected it and regenerated true b_{ij}s (both reference and focal) for item i until an ordered set of b_{ijF}s was obtained.

Procedures

Each data set was analyzed four times using source code from Thissen's IRTLRDIF program, Version 2.0b (2001). The four analyses differed with respect to the number of anchors. First, every item was tested for DIF with all others as anchors. Then designated anchors were selected based on the magnitude of the LR/f ratio, as described above. The number of designated anchors was either 1, 10% of k, or 20% of k. Next, all items not selected to be anchors were tested for DIF. Data generated for 10-item tests were analyzed three rather than four times because 10% of 10 is 1.

In every analysis, the θ distribution was assumed standard normal for the reference group and normal for the focal group, with the mean and standard deviation estimated simultaneously with the item parameters. If a zero cell frequency was generated (i.e., no simulees responded in a particular category to an item), the categories for that item were collapsed and estimation proceeded as usual with one fewer b_{ij} parameter. The latent variable was represented with rectangular quadrature, ranging from -6 to 6 in increments of 0.1 (121 points). The maximum number of expectation-maximization (EM) cycles was 2,000 for models with item parameters constrained equal between groups and 1,000 for models with item parameters permitted to vary between groups. Pilot research showed that maximum numbers of EM cycles half as large would have sufficed for multiple-anchor models but were inadequate for single-anchor models. A fitting was declared converged when the parameter estimate that was changing the most between cycles changed less than .0001.

The Benjamini-Hochberg (BH; Benjamini & Hochberg, 1995) procedure has been recommended for controlling the false discovery rate for IRT-LRT (Thissen, Steinberg, & Kuang, 2002; Williams, Jones, & Tukey, 1999). With the BH method, p values for tests $i = 1, 2, \ldots, m$ are ordered from smallest to largest, and then test i is declared significant if p_i fails to exceed the critical value

$$\frac{i\alpha}{2m},$$

where α is the Type I error rate and m is the total number of tests. In the present research, the BH adjustment described in the documentation for the SAS MULTTEST procedure (Version 9.1) was implemented in C ++ and applied separately for each simulated data set ($\alpha = .05$).

Table 1
Patterns of Differential Item Functioning in Simulated Threshold Parameters

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

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

Outcomes

Proportion of invariant anchors. The proportion of replications with a group-invariant ("clean") anchor set, versus an anchor set contaminated by one or more DF items ("dirty"), was tabulated for each condition and number of anchors.

Accuracy of the hypothesis tests. Four variables quantified accuracy of the DIF tests: (a) mean number of studied items with significant tests (averaged over replications and number of anchors), (b) number of replications for which results produced the "correct" model, (c) hit rate, and (d) false alarm rate. Results produced the correct model when the anchor set was clean, all DF items had significant tests, and all group-invariant studied items had nonsignificant tests. The hit rate was the proportion of DF studied items for which the DIF test was significant, and the false alarm rate was the proportion of group-invariant studied items for which the DIF test was significant. The rates were averaged over clean and dirty replications separately for each condition and number of anchors.

Average unsigned difference (AUD). To judge recovery of the magnitude of DIF, an effect size was computed for each studied item that was averaged separately over items that did or did not function differently between groups for each replication, then averaged separately over clean and dirty replications for each condition and number of anchors.

The AUD between the expected response functions (ERFs) for the focal and reference groups (Wainer, 1993) was used to measure the magnitude of DIF for each studied item. An ERF gives the item score (e.g., 1, 2, 3, 4, or 5) expected at each value of θ (D. M. Bolt et al., 2004; Steinberg & Thissen, 2006; Wainer, Sireci, & Thissen, 1991). Here, ERFs were computed with θ represented by 121 quadrature points between −6 and 6 in 0.1 increments. The AUD was calculated as the absolute value of the difference between the two ERFs at each quadrature point, weighted by the density of the normal focal-group θ distribution at that point, averaged over points.

One true AUD and several estimated AUDs were computed for each simulation condition. The true AUD used true item parameters and the true mean and standard deviation of θ for the focal group (−0.4 and 1, respectively). There was an estimated AUD for each separate anchor set, calculated using item parameters and the focal-group mean and standard deviation estimated from the model that permitted the studied item's parameters to vary between groups. To control outliers, any a_i estimates greater than 4 were recoded to 4 before the ERFs were computed.

Estimated mean and standard deviation of θ for the focal group. In IRT-LRT, the focal-group mean and standard deviation of θ were estimated simultaneously with the item parameters separately for every studied item from the model permitting the studied item's parameters to vary between groups. Here, the mean and standard deviation estimates were averaged over all studied items for each replication, then averaged over clean and dirty replications separately for each condition and anchor set. Because the mean of θ was fixed to be 0 for the reference group when the model was fitted, the absolute value of the focal-group mean was the estimated group-mean difference.

Results

Convergence

Convergence rates were high, but were least high for fittings using a single anchor. Convergence rates decreased with increases in the percentage of DF items. With all others as anchors, all models in all 21 conditions converged. Convergence rates ranged from 99.4% to 100% with .2*k* anchors, 99.6% to 100% with .1*k* anchors, and 97.3% to 100% with 1 anchor. All results were analyzed, regardless of convergence, because (a) there were very few nonconverged models, (b) nonconvergence was about equally likely for items that did versus did not function differently between groups, and (c) the mean LR statistic was nearly identical for converged versus nonconverged fittings.

Large Discrimination Parameter Estimates ($a_i > 4$)

The number of large a_i estimates was greater for the focal group and increased with increases in the percentage of DF items and decreases in the number of anchors. The percentage of large a_is ranged from 0% to 1.2% with all others as anchors, 0% to 11.8% with .2*k* anchors, .02% to 10.9% with .1*k* anchors, and 1.1% to 71.7% with 1 anchor.

Table 2: 0% or 20% of Items Function Differently

The proposed method for empirically selecting anchors produced a group-invariant (clean) set of designated anchors in all replications in all conditions when 20% of items functioned differently (one exception: in the condition with $k = 40$, eight anchors, and uniform DIF, one replication had a dirty anchor). With a clean anchor set, IRT-LRT performed well; however, the number of correct models was smaller for data with DIF than for DIF-free data.

Table 2 shows that almost all results were quite accurate with any number of designated anchors (1, .1*k*, or .2*k*), less accurate with all others as anchors, and usually best with $1 < g <$ all others. When 20% of items functioned differently, the number of correct models was always higher with designated anchors than with all others as anchors and usually greater with .2*k* anchors (10-item tests) or .1*k* anchors (20- or 40-item tests) than with a single anchor. Hit rates were near 1 with multiple anchors but still high (.90 to .98) with one anchor. False alarm rates increased with

Table 2
Simulation Results: 0% or 20% of Items Function Differently

	10 Items			20 Items			40 Items		
Items With DIF	0	2		0	4		0	8	
Type of DIF		b_{ij}	a_i & b_{ij}		b_{ij}	a_i & b_{ij}		b_{ij}	a_i & b_{ij}
				Single anchor					
Hit rate	—	.94	.94	—	.92	.97	—	.90	.98
False alarm rate	.00	.00	.01	.00	.01	.00	0	.00	.01
Mean # sig. tests	0.01	1.91	1.93	0.01	3.84	3.92	0	7.22	7.97
$\bar{\theta}_F$	–.42	–.38	–.36	–.40	–.37	–.37	–.41	–.38	–.36
SD_{θ_F}	1.01	0.99	0.96	1.00	0.97	0.97	1.01	0.98	0.95
AUD (DIF)	—	.06	.06	—	.06	.06	—	.06	.06
AUD (invariant)	.01	.01	.01	.01	.01	.01	.01	.01	.01
Correct models	99	85	85	99	63	87	100	46	71
				Anchor: 10% of k					
Hit rate				—	.99	1	—	.98	1
False alarm rate				.00	.02	.01	.00	.02	.02
Mean # sig. tests				0.05	4.23	4.06	0.02	8.30	8.50
$\bar{\theta}_F$		(Above)		–.41	–.36	–.36	–.41	–.37	–.36
SD_{θ_F}				1.01	0.97	0.98	1.00	0.98	0.97
AUD (DIF)				—	.05	.06	—	.05	.06
AUD (invariant)				.01	.02	.01	.01	.01	.01
Correct models				95	73	93	98	53	61
				Anchor: 20% of k					
Hit rate	—	.98	1	—	1	1	—	.99	1
False alarm rate	.01	.02	.02	.00	.03	.02	.00	.03	.03
Mean # sig. tests	0.04	2.06	2.10	0.05	4.36	4.21	0.03	8.59	8.73
$\bar{\theta}_F$	–.41	–.36	–.36	–.40	–.36	–.37	–.40	–.38	–.37
SD_{θ_F}	1.01	0.98	0.98	1.01	0.97	0.98	1.00	0.99	0.98
AUD (DIF)	—	.05	.06	—	.05	.06	—	.05	.06
AUD (invariant)	.01	.02	.01	.01	.01	.01	.01	.01	.01
Correct models	96	88	93	95	69	79	97	47	51
				All-others anchor					
Hit rate	—	.99	1	—	1	1	—	.99	1
False alarm rate	.00	.12	.17	.00	.09	.08	.00	.06	.07
Mean # sig. tests	0.03	2.93	3.36	0.07	5.42	5.35	0.03	9.98	10.20
$\bar{\theta}_F$	–.41	–.33	–.33	–.40	–.33	–.34	–.41	–.35	–.34
SD_{θ_F}	1.01	0.96	0.95	1.00	0.96	0.96	1.01	0.97	0.95
AUD (DIF)	—	.05	.06	—	.05	.06	—	.05	.06
AUD (invariant)	.01	.02	.01	.01	.01	.01	.01	.01	.01
Correct models	97	49	44	93	39	39	97	27	27

Note. DIF = differential item functioning; mean # sig. tests = mean number of items with significant DIF tests; $\bar{\theta}_F$ = focal group mean; SD_{θ_F} = focal group standard deviation; AUD = average unsigned difference between expected response functions; k = total items; — = not applicable.

decreasing k and were near 0 with designated anchors (.00 to .03) but larger (.06 to .17) with all others as anchors. Consistently, the mean number of items with significant tests was near the true number with each set of designated anchors but elevated with all others as anchors. Although the mean and standard deviation of θ for the focal group were misestimated in all conditions (except those without DIF), the bias was small.

With designated anchors, power tended to be greater with nonuniform versus uniform DIF. This appeared to be due to the simulation methodology: True a_is tended to be greater in conditions with nonuniform DIF because the distribution from which they were drawn was truncated at 1.2 instead of 0.5. The lower limit for a_{iR} was different for items with versus without DIF in a_i to ensure that $a_{iF} \geq .5$ (the largest possible amount of DIF in a_i was .7). This power difference probably explains why the number of correct models was usually greater for conditions with nonuniform DIF (Tables 2, 3, and 4).

The AUD was not influenced by the number of anchors. For each condition and anchor set (including all others as anchors), the mean AUD was near the true value of 0 for group-invariant items, .05 for items with uniform DIF, or .06 for items with nonuniform DIF.

Table 3: 50% of Items Function Differently

When 50% of the items functioned differently, the anchor set selected by the rank-based method was clean for most replications. Across conditions, the percentage of clean anchor sets was 98% to 100% with a single anchor, 100% with .1k anchors, and 96% to 100% with .2k anchors.

Results in Table 3 show that IRT-LRT performed well with a clean anchor set, better with any number of clean anchors than with all others as anchors, and usually best with $1 < g <$ all others. With all others as anchors, hit rates were high, but all other outcomes were very erroneous. Although results were highly variable and sometimes quite poor with a dirty designated anchor set, most designated anchors were clean. For each condition, mean AUDs (not shown) for clean anchor sets were nearly identical to those in Table 2. For dirty anchor sets, mean AUDs tended to be about .04 for all items regardless of DIF status.

Table 4: 80% of Items Function Differently

When 80% of items functioned differently, the number of replications with a clean anchor set depended on the number of anchors: The more items chosen, the greater the chance of contamination. If only one anchor was selected, that item was clean between 71% and 96% of the time. However, with multiple designated anchors, a clean set was chosen for 57% to 83% of replications for .1k anchors and 20% to 56% for .2k anchors. Notice that when 80% of items functioned differently, .2k items were invariant; thus, when the set of .2k anchors was clean, none of the studied items were invariant and false alarm rates (and AUDs for invariant items) could not be computed.

Results in Table 4 show that IRT-LRT performed well with a clean anchor set, better with any number of clean designated anchors than with all others as anchors, and usually best with $1 < g <$ all others. For each condition, mean AUDs (not shown) for clean anchor sets were nearly identical to those in Table 2. For dirty anchor sets, mean AUDs ranged from .03 to .06 in the presence of DIF and .02 to .09 in the absence of DIF.

Outcomes (besides hit rates) were extremely inaccurate with all others as anchors. However, when a designated anchor set was contaminated, false discovery rates, estimates of the focal-group θ parameters, and AUDs were also poor and sometimes worse than those using all others as anchors. For example, with nonuniform DIF in 80% of items, the false alarm rate in dirty replications was .93

Table 3
Simulation Results: 50% of Items Function Differently

	10 Items				20 Items				40 Items			
Type	b_{ij} Only		a_i & b_{ij}		b_{ij} Only		a_i & b_{ij}		b_{ij} Only		a_i & b_{ij}	
Single anchor												
# Sig.	4.68		4.95		9.25		9.90		18.00		19.70	
Correct	59		79		41		75		20		43	
Anchor	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty
Reps	99	1	98	2	100	0	100	0	100	0	100	0
Hit	.91	.75	.97	.25	.91	—	.98	—	.89	—	.97	—
False	.03	.20	.02	1	.02	—	.01	—	.01	—	.02	—
$\bar{\theta}_F$	–.38	.10	–.35	.29	–.37	—	–.33	—	–.35	—	–.31	—
SD_{θ_F}	0.95	1.06	0.90	0.74	0.95	—	0.90	—	0.96	—	0.88	—
Anchor: 10% of k												
# Sig.					10.20		10.20		20.70		21.00	
Correct					65		75		44		44	
Anchor					Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty
Reps					100	0	100	0	100	0	100	0
Hit	(Above)				.99	—	1	—	.99	—	1	—
False					.05	—	.03	—	.05	—	.06	—
$\bar{\theta}_F$					–.35	—	–.34	—	–.33	—	–.33	—
SD_{θ_F}					0.95	—	0.97	—	0.96	—	0.95	—
Anchor: 20% of k												
# Sig.	5.15		5.26		10.30		10.30		20.70		21.00	
Correct	72		83		58		75		38		43	
Anchor	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty
Reps	98	2	96	4	99	1	100	0	98	2	98	2
Hit	.98	1	1	1	.99	1	1	—	1	1	1	1
False	.08	.25	.05	1	.07	.43	.05	—	.07	.12	.08	.19
$\bar{\theta}_F$	–.37	–.14	–.37	.06	–.36	–.31	–.35	—	–.35	–.45	–.35	–.40
SD_{θ_F}	0.96	0.89	0.97	0.70	0.97	1.11	0.97	—	0.98	1.03	0.97	0.92
All-others anchor												
# Sig.	7.52		7.76		14.00		15.70		27.50		30.10	
Correct	14		10		4		2		1		1	
Anchor	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty
Hit	—	.99	—	1	—	1	—	1	—	1	—	1
False	—	.51	—	.55	—	.41	—	.57	—	.38	—	.51
$\bar{\theta}_F$	—	–.24	—	–.23	—	–.26	—	–.21	—	–.26	—	–.24
SD_{θ_F}	—	0.90	—	0.87	—	0.93	—	0.89	—	0.94	—	0.89

Note. # Sig. = mean number of items with significant differential item functioning tests; correct = number of correct models; reps = number of replications; $\bar{\theta}_F =$ mean of θ for the focal group; $SD_{\theta_F} =$ standard deviation of θ for the focal group; — = not applicable.

Table 4
Simulation Results: 80% of Items Function Differently

	10 Items				20 Items				40 Items			
Type	b_{ij} Only		a_i & b_{ij}		b_{ij} Only		a_i & b_{ij}		b_{ij} Only		a_i & b_{ij}	
Single anchor												
# Sig.	7.43		7.81		14.60		15.70		29.30		31.10	
Correct	49		55		34		66		7		42	
Anchor	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty
Reps	87	13	71	29	92	8	87	13	96	4	93	7
Hit	.93	.86	.97	.86	.92	.70	.99	.72	.91	.90	.97	.64
False	.03	.69	.01	.93	.01	.47	.02	1	.02	.06	.03	.96
$\bar{\theta}_F$	− .41	.07	− .36	.26	− .38	.00	− .36	.28	− .37	− .16	− .31	.36
SD_{θ_F}	0.96	0.94	0.90	0.59	0.96	0.89	0.87	0.60	0.94	0.90	0.86	0.66
Anchor: 10% of k												
# Sig.					16.00		16.30		32.30		33.20	
Correct					57		73		41		42	
Anchor					Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty
Reps	(Above)				83	17	81	19	67	33	57	43
Hit					.98	.99	1	.93	.99	1	1	1
False					.06	.60	.05	.98	.09	.35	.08	.71
$\bar{\theta}_F$					− .35	− .18	− .39	.14	− .36	− .26	− .37	− .10
SD_{θ_F}					0.99	0.95	0.97	0.71	0.98	0.97	0.98	0.84
Anchor: 20% of k												
# Sig.	7.84		7.91		15.50		15.80		31.00		31.70	
Correct	54		52		38		51		19		20	
Anchor	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty
Reps	56	44	52	48	43	57	51	49	20	80	20	80
Hit	1	1	1	.99	.99	.99	1	1	1	1	1	1
False	—	.68	—	.90	—	.38	—	.67	—	.26	—	.72
$\bar{\theta}_F$	− .40	− .13	− .39	.04	− .38	− .28	− .42	− .12	− .40	− .33	− .40	− .17
SD_{θ_F}	1.00	0.89	0.98	0.74	1.00	0.99	0.99	0.85	1.00	0.99	1.00	0.88
All-others anchor												
# Sig.	9.52		9.66		18.90		19.30		37.70		38.80	
Correct	9		7		3		1		2		0	
Anchor	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty	Clean	Dirty
Hit	—	1	—	1	—	1	—	1	—	1	—	1
False	—	.77	—	.84	—	.75	—	.83	—	.72	—	.85
$\bar{\theta}_F$	—	− .15	—	− .10	—	− .15	—	− .13	—	− .17	—	− .11
SD_{θ_F}	—	0.85	—	0.83	—	0.89	—	0.83	—	0.90	—	0.84

Note. # Sig. = mean number of items with significant differential item functioning tests; correct = number of correct models; reps = number of replications; $\bar{\theta}_F$ = mean of θ for the focal group; SD_{θ_F} = standard deviation of θ for the focal group; — = not applicable.

with a single anchor versus .84 with all others as anchors ($k = 10$), and .98 with $.1k$ anchors versus .83 with all others as anchors ($k = 20$).

Sensitivity Analysis: Smaller *N*

Some researchers do not have access to samples as large as those simulated here; thus, it is useful to examine results with a smaller sample. Two simulation conditions were repeated with $N_R = 600$, $N_F = 200$: (a) $k = 20$, 20% with nonuniform DIF, and (b) $k = 10$, 50% with uniform DIF. All other simulation and analytic procedures described above were replicated.

Convergence was nearly perfect (99.9% to 100% of all models converged). The proportion of large a_is was in the same range and depended on the same variables as with larger N. All empirically selected anchors were clean for all replications for $k = 20$ (20% DIF). For $k = 10$ (50% DIF), single-anchor and two-anchor sets were clean for 98 and 95 replications, respectively.

The most important finding with smaller N was lower power to detect DIF. Hit rates, the mean number of significant tests, and the number of correct models were lower with smaller N. However, N interacted with the number of anchors. With smaller N and a single invariant anchor, fewer correct models were identified than with all others as anchors. This was apparently due to low power for single anchors: Hit rates were .75 ($k = 20$) and .71 ($k = 10$, clean replications).

Results were better with multiple clean anchors. With $k = 20$ (20% DIF), hit rates were .96 (two anchors) and .99 (four anchors), and the number of correct models was 79 (two or four anchors), versus 60 (all others as anchors) and 40 (single anchor). With $k = 10$ (50% DIF), the hit rate was .92 (two anchors), and the number of correct models was 57 (two anchors) versus 27 (all others) and 25 (single anchor). Complete results with smaller N are available by request.

Discussion

A quick and easy rank-based strategy for empirically selecting designated anchors was proposed. Simulation results supported the utility of the strategy for IRT-LRT carried out as implemented in IRTLRDIF (Thissen, 2001). The rank-based method almost always produced a DIF-free anchor set when 20% or 50% of items functioned differently, even when the anchor set consisted of 20% of the total items. When the anchor set was invariant, results were quite accurate and much better than those with all others as anchors.

It was more difficult to select clean anchor sets when 80% of items functioned differently, probably because the LR statistic reflects model fit as a function of the parameter estimates, which are less accurate with increasing DIF in the data. When the anchor set was contaminated, the false discovery rate and estimates of the focal-group θ parameters and AUDs were inaccurate—sometimes even more so than with all others as anchors. It is hard to imagine an anchor-selection method for which performance would not decline with increasing DIF. The challenge for future research is to find the method that is least inaccurate under these suboptimal conditions. A study comparing the strategy proposed here to the various other suggestions for empirically selecting anchors is needed.

Use of a single anchor will minimize the chance of contamination, which is especially important when the percentage of DF items is large. Estimation is less stable and power is lower with a single anchor, but present and previous (Stark et al., 2006; Wang, 2004; Wang & Yeh, 2003) simulations show that results can be quite accurate with larger N. However, with smaller N, power was so low with a single anchor that the chance of finding the correct model was worse than with all others as anchors. This was true with 20% or 50% DF items. Therefore, single anchors are not recommended with smaller samples.

There may be a conceptual disadvantage of a single anchor that has not been addressed in simulations. The anchor set defines the matching variable, θ, which may not be very well characterized by a single item. Thus, the meaning of θ may be different than intended with a single anchor. Edelen et al. (2006) and Williams (1997) mention that a good anchor set has high discrimination ability with threshold parameters spread over a relatively wide range of θ. These criteria ensure that the construct is well defined. Decisions about the number of anchors must balance the need to avoid contamination with the need for high validity. Balancing these needs is easy with relatively few DF items but becomes increasingly difficult as the percentage of DF items increases. Hopefully, it is rare in practice for as many as 80% of items to function differently.

One way to increase validity is to use IRT-LRT only for decisions about the presence or absence of significant DIF and then to fit a final model for estimates of the item parameters and the focal-group θ parameters. In the final model, one set of parameters is estimated for all invariant items, and group-specific parameters are estimated for DF items. With all items included in the likelihood and more items linking the metric between groups, the meaning of θ and the parameter estimates are likely to be more valid. Also, the focal-group mean and standard deviation from the final model should be more accurate than those estimated from the models used to test individual items for DIF. This simulation study did not examine results from final models, but it would be informative to evaluate them in the future.

This research leads to some specific recommendations for practitioners of IRT-LRT. The recommended steps are to

1. carry out IRT-LRT with all others as anchors;
2. select g items with the g smallest LR/f ratios as designated anchors;
3. test each studied item for DIF using the designated anchors; and
4. based on the DIF results, fit a multiple-group model using a program such as MULTILOG (Thissen, 1991) to obtain final parameter estimates.

Anchors need to be valid indicators of the intended construct, and g should usually be approximately 10% to 20% of the total number of items. If many items (e.g., 80% or more) have large LR/f ratios in the initial application of IRT-LRT (with all others as anchors), it is probably better to designate only a single anchor to decrease the chance of contamination. Statistical power should be adequate with a single anchor if the sample size is relatively large (e.g., $N_R = 1,500$, $N_F = 500$). However, if the sample size is smaller (e.g., $N_R = 600$, $N_F = 200$), g must exceed 1 to ensure adequate power. Unless designated anchors can be well chosen based on extensive previous research, IRT-LRT is not recommended when the sample is small and the number of DF items appears to be large.

Finally, the rank-based strategy is not specific to IRT-LRT and could be used with, for example, MH, logistic regression, or SIBTEST (Shealy & Stout, 1993) methods for testing individual items for DIF. With these methods, reference and focal group members are matched on summed scores, and the inclusion of all items in the summed score is analogous to the use of all others as anchors in IRT-LRT. To use the rank-based strategy, (a) compute test statistics with total scores as the matching criterion, (b) select g items with the g smallest test statistics as designated anchors, and (c) test each studied item for DIF using the designated anchors (plus the studied item; see Holland & Thayer, 1988) for the matching criterion. It would be useful to empirically evaluate the rank-based strategy for DIF-testing methods other than IRT-LRT.

References

*References marked with an asterisk contain the applications that were used to guide the simulation details.

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.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289-300.

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

*Bolt, D. 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.

*Bolt, S. E., & Ysseldyke, J. E. (2006). Comparing DIF across math and reading/language arts tests for students receiving a read-aloud accommodation. *Applied Measurement in Education*, *19*, 329-355.

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

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253-260.

*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., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research and Practice*, *20*, 225-233.

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.

*Cooke, D. J., & Michie, C. (1999). Psychopathology across cultures: North America and Scotland compared. *Journal of Abnormal Psychology*, *108*, 58-68.

*Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach. *Medical Care*, *44*, S134-S142.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*, 278-295.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics*, *5*, 1148-1169.

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

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

*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, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*, 89-114.

*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.

Kok, F. G., Mellenbergh, G. J., & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement*, *22*, 295-303.

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

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2(dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*, 55-64.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361-388.

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

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

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

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

*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.

*Pae, T. (2004). DIF for examinees with different academic backgrounds. *Language Testing*, *21*, 53-73.

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

*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.

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Reider (Eds.), *Equivalence in measurement: Research in management* (pp. 25-50). Greenwich, CT: Information Age.

*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. (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: Springer.

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

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. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*, *66*, 341-349.

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

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

Thissen, D. (1991). MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory [Computer software and manual]. Chicago, IL: 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. Documentation for computer program [Computer software and manual]. 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., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*, 77-83.

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: 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. 67-111). Hillsdale, NJ: Lawrence Erlbaum.

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

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

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.

*Williams, V. S. L. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education*, *10*, 253-267.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, *24*, 42-69.

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.