A Comparison of Lord's χ^2 and Raju's Area Measures In Detection of DIF

Allan S. Cohen and Seock-Ho Kim

University of Wisconsin

The area between item response functions estimated in different samples is often used as a measure of differential item functioning (DIF). Under item response theory, this area should be 0, except for errors of measurement. This study examined the effectiveness of two statistical tests of this area—a Z test for exact signed area and a Z test for exact unsigned area—for different test length, sample size, proportion of DIF items on the test, and item parameter estimation conditions using the two-parameter model. Errors in detection made using these two statistics were compared with errors made using Lord's χ^2. Differences between all three statistics were relatively small; however, the χ^2 statistic was more effective than either of the two Z tests at detecting simulated DIF. The Z test for the exact signed area was the least effective and was the most likely to result in false negative errors. *Index terms: area measures, differential item functioning, item response theory, item bias, Lord's* χ^2.

When the same test item is administered to different samples of examinees, one measure of differential item functioning (DIF) in the item is whether the item parameters are equal. Under item response theory (IRT), the item response function (IRF) defined by the parameters for the item estimated in different groups should be the same within a linear transformation. One measure of this equivalence is provided by Lord's (1980) χ^2 statistic, which tests the hypothesis that each of the parameters of the IRF are identical across groups.

In a typical DIF study, the item parameters in one group—the focal group (F)—are compared to those estimated in a second group—the reference group (R). In the two-parameter model (2PM), the probability of examinee i giving a correct response to item j is a function of the discrimination of the item, a_j, the difficulty of the item, b_j, and the unidimensional trait of the examinee, θ_i. This probability is expressed as

$$P_j(\theta_i) = \{1 + \exp[-Da_j(\theta_i - b_j)]\}^{-1} , \tag{1}$$

where D is a scaling constant equal to 1 for the logistic model and 1.702 to approximate the normal ogive model from the logistic model.

For the 2PM, Lord's (1980) χ^2 method simultaneously tests the null hypothesis, H_o: $a_{jR} = a_{jF}$ and $b_{jR} = b_{jF}$ for item j, where a_{jR} and a_{jF} are the discrimination parameters, and b_{jR} and b_{jF} are the difficulty parameters of item j in the reference and focal groups, respectively. The χ^2 statistic is defined in terms of parameter estimates, $\hat{a}_j$ and $\hat{b}_j$, as

$$\chi_j^2 = (\xi_{jR} - \xi_{jF})'\Sigma_j^{-1} (\xi_{jR} - \xi_{jF}) , \tag{2}$$

where
$\xi_{jR} = (\hat{a}_{jR}\ \hat{b}_{jR})'$,
$\xi_{jF} = (\hat{a}_{jF}\ \hat{b}_{jF})'$, and

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 1, March 1993, pp. 39-52
© Copyright 1993 Applied Psychological Measurement Inc.
0146-6216/93/010039-14$1.95

$\boldsymbol{\Sigma}_j$ is the 2×2 dispersion matrix that has the form

$$\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_{j\mathrm{R}} + \boldsymbol{\Sigma}_{j\mathrm{F}} \ , \tag{3}$$

where $\boldsymbol{\Sigma}_{j\mathrm{R}}$ and $\boldsymbol{\Sigma}_{j\mathrm{F}}$ are the variance and covariance matrices of $\xi_{j\mathrm{R}}$ and $\xi_{j\mathrm{F}}$, respectively.

Lord's χ^2 has been used in a number of studies (e.g., Candell & Hulin, 1987; McCauley & Mendoza, 1985) and has been shown to be relatively useful for detecting DIF. An important assumption for the use of this statistic, however, is that θ is known (Lord, 1980). McLaughlin & Drasgow (1987) showed that, when both θ and item parameters are unknown, the probability of making a Type I error may be substantial.

A similar approach in detecting DIF has been to estimate the area between two IRFs. There is some evidence that area measures can provide useful information regarding the presence of DIF (Ironson & Subkoviak, 1979; Linn, Levine, Hastings, & Wardrop, 1981; McCauley & Mendoza, 1985; Rudner, Getson, & Knight, 1980; Shepard, Camilli, & Williams, 1984, 1985). Under IRT, if this area is nonzero, the item is functioning differentially in the two groups. The two area measures most frequently used—the signed area and the unsigned area—can be measured over either a bounded (closed) interval or an open (exact) interval on the θ scale.

A serious drawback to using either a closed-interval (Kim & Cohen, 1991) or an exact-interval (Raju, 1988) measure, however, has been lack of a statistical test of the area between two IRFs. Although no such tests are as yet available for the closed-interval measures, Raju (1990) presented information regarding the sampling distributions for the two exact area measures and proposed significance tests for each. Two statistics were proposed: *Z(ESA)* and *Z(H)* for the exact signed area (*ESA*) and exact unsigned area (*EUA*), respectively.

For the 2PM, the *ESA* and *EUA* for item j are defined by Raju (1990) as

$$ESA = \hat{b}_{j\mathrm{F}} - \hat{b}_{j\mathrm{R}} \tag{4}$$

and

$$EUA = \begin{cases} |\hat{b}_{j\mathrm{F}} - \hat{b}_{j\mathrm{R}}| & \text{if } \hat{a}_{j\mathrm{R}} = \hat{a}_{j\mathrm{F}} \\ |H_j| & \text{otherwise} \end{cases} \ , \tag{5}$$

where

$$H_j = \frac{2(\hat{a}_{j\mathrm{F}} - \hat{a}_{j\mathrm{R}})}{D\hat{a}_{j\mathrm{R}}\hat{a}_{j\mathrm{F}}} \ln\left\{1 + \exp\left[\frac{D\hat{a}_{j\mathrm{R}}\hat{a}_{j\mathrm{F}}(\hat{b}_{j\mathrm{F}} - \hat{b}_{j\mathrm{R}})}{\hat{a}_{j\mathrm{F}} - \hat{a}_{j\mathrm{R}}}\right]\right\} - (\hat{b}_{j\mathrm{F}} - \hat{b}_{j\mathrm{R}}) \ . \tag{6}$$

The test statistic, Z_j, for *ESA* is defined as

$$Z_j(ESA) = \frac{\hat{b}_{j\mathrm{F}} - \hat{b}_{j\mathrm{R}}}{[\mathrm{Var}(\hat{b}_{j\mathrm{F}}) + \mathrm{Var}(\hat{b}_{j\mathrm{R}})]^{1/2}} \ , \tag{7}$$

where the asymptotic variances of $\hat{b}_{j\mathrm{R}}$ and $\hat{b}_{j\mathrm{F}}$ can be obtained in a comparable manner to that in Equation 26 in Raju (1990).

Because asymptotic normality cannot be assumed for the *EUA*, Raju (1990) suggested that H should be used to test the significance of *EUA* (or *ESA* may be used when $\hat{a}_{j\mathrm{R}} = \hat{a}_{j\mathrm{F}}$). The test statistic Z_j for H is

$$Z_j(H) = \frac{H_j}{[\mathrm{Var}(H_j)]^{1/2}} \ , \tag{8}$$

where the variance for H_j can be obtained in a manner similar to that in Equation 34 in Raju (1990).

These significance tests, if effective, could represent important contributions to the study of DIF, but the extent to which *Z(ESA)* and *Z(H)* are able to detect DIF items has not been studied. Therefore, the present study was designed to compare the effectiveness of Raju's *Z(ESA)* and *Z(H)* to Lord's χ^2 on simulated datasets.

Method

Data Generation

A 2PM was used to generate the simulated datasets using the computer program GENIRV (Baker, 1986). Item and θ parameters were expressed in the logistic metric (i.e., $D = 1$).

Two test lengths were simulated—20 and 60 items. In addition, two different sample sizes were simulated—100 and 500 examinees. Test length and sample size were completely crossed to yield four test length × sample size conditions.

Data were generated for one reference group and two different types of focal groups based on their underlying θ distributions. The matched-θ focal groups were matched on θ with the reference group. The nonmatched-θ focal groups had mean θs of 1 standard deviation below the reference group. Generating parameters for the underlying θ distributions for the reference and the matched-θ focal groups were normal (0, 1). The generating θ distributions for the nonmatched-θ focal groups were normal (–1, 1).

Three different DIF conditions were generated for each of the focal groups: a null or 0% DIF condition, a 10% DIF condition, and a 20% DIF condition. For the 0% DIF condition, all item parameters were the same as those in the reference group. For the 10% DIF condition, 10% of the item parameters were changed, and for the 20% DIF condition, 20% of the item parameters were changed. The generating parameters for the 20-item tests are given in Table 1. Generating parameters similar to those in the 20-item tests were used three times in the 60-item tests. The generating parameters for the 60-item tests are given in Table 2. Note that the locations of the repeated items in the 60-item tests were shifted to make them contiguous to the first occurrence for each item.

Five replications were conducted for each simulated condition. This yielded 140 different datasets: two test lengths, two sample sizes, seven groups (one reference group, three matched-θ focal groups, and three nonmatched-θ focal groups), and five replications. Parameters in each dataset were estimated using two different estimation algorithms (see below). Six reference group and focal group comparisons were formed by matching item parameters obtained from the reference group and one of the six different focal groups generated for each test length × sample size condition. The three matched-θ comparisons were formed as follows: the 0-Matched group consisted of the reference group compared to the 0% DIF matched-θ focal group, the 10-Matched group consisted of the reference group compared to the 10% DIF matched-θ focal group, and the 20-Matched group consisted of the reference group compared to the 20% DIF matched-θ focal group. Three similar nonmatched-θ comparisons were formed for the three DIF conditions. The three nonmatched-θ comparisons were formed as follows: The 0-Nonmatched group consisted of the reference group compared to the 0% DIF nonmatched-θ focal group, the 10-Nonmatched group consisted of the reference group compared to the 10% DIF nonmatched-θ focal group, and the 20-Nonmatched group consisted of the reference group compared to the 20% DIF nonmatched-θ focal group. DIF detection was studied only for comparisons of reference and focal groups of equal sample size within each test length × sample size condition.

Table 1
Item Parameters Used to Generate the 20-Item Datasets for the Reference and 0-Focal Groups, the 10-Focal Group [Containing Two Uniform DIF Items (5,10)], and the 20-Focal Group [Containing Two Uniform DIF Items (5,10) and Two Nonuniform DIF Items (15,20)] (Blanks Indicate That the Same Item Parameters Were Used)

	Reference and 0-Focal		10-Focal		20-Focal	
Item	*a*	*b*	*a*	*b*	*a*	*b*
1	.55	0.00				
2	.73	-1.04				
3-4	.73	0.00				
5	.73	1.04	.73	2.04	.73	2.04
6	1.00	-1.96				
7-8	1.00	-1.04				
9	1.00	0.00				
10	1.00	0.00	1.00	.50	1.00	.50
11-12	1.00	0.00				
13-14	1.00	1.04				
15	1.00	1.96			.50	2.46
16	1.36	-1.04				
17-18	1.36	0.00				
19	1.36	1.04				
20	1.80	0.00			1.30	0.00

Parameter Estimation

Item and θ parameters were estimated using BILOG 3 (Mislevy & Bock, 1990). BILOG 3 default conditions implement a marginal Bayesian estimation (MBE; i.e., marginal maximum a posteriori estimation) procedure for the 2PM. Previous research using area measures has suggested that the algorithm used to estimate item parameters may influence the detection of DIF (Cohen, Kim, & Subkoviak, 1991). Because these parameter estimates are the basis for Lord's χ^2 and Raju's significance tests, it is possible that similar effects might be present for these tests as well. Therefore, item and θ parameters also were estimated using marginal maximum likelihood estimation (MMLE). Under MMLE, a maximum likelihood estimation algorithm was used to estimate θ. For MBE, the default prior in BILOG 3, which is on the item discrimination, was used and a Bayesian expected a posteriori estimation algorithm was used for θ.

Linking of Metrics

Linking method. IRT DIF studies require that the item parameter estimates for the same item from different samples of examinees first be expressed in the same metric. The transformation or linking of the metric obtained in the focal group to that of the reference group can be affected by the presence of DIF items to the extent that errors in the linking can result in errors in detection of DIF items (Shepard et al., 1984).

Recent evidence (Kim & Cohen, 1992) suggests that the test characteristic curve (or test response function, TRF) method for linking (Stocking & Lord, 1983) is more accurate for small samples than either a weighted mean and sigma method (Linn et al., 1981) or the minimum χ^2 method (Divgi, 1985). Because one objective of the present study was to investigate DIF detection with small sample sizes, the TRF linking method was used for all conditions as implemented by Baker, Al-Karni, & Al-Dosary (1991).

Table 2
Item Parameters Used to Generate the 60-Item Datasets for the Reference and 0-Focal Groups, the 10-Focal Group [Containing Six Uniform DIF Items (5, 10, 15, 20, 25, 30)] and the 20-Focal Group [Containing Six Uniform DIF Items (5, 10, 25, 30, 45, 50) and Six Nonuniform DIF Items (15, 20, 35, 40, 55, 60)] (Blanks Indicate That the Same Item Parameters Were Used)

	Reference and 0-Focal		10-Focal		20-Focal	
Item	*a*	*b*	*a*	*b*	*a*	*b*
1-3	.55	0.00				
4	.73	-1.04				
5	.73	-1.04	.73	-.04	.73	-.04
6	.73	-1.04				
7-9	.73	0.00				
10	.73	0.00	.73	.50	.73	.50
11-12	.73	0.00				
13-14	.73	1.04				
15	.73	1.04	.73	2.04	.23	1.54
16-18	1.00	-1.96				
19	1.00	-1.04				
20	1.00	-1.04	1.00	-.54	.50	-1.04
21-24	1.00	-1.04				
25	1.00	0.00	1.00	1.00	1.00	1.00
26-29	1.00	0.00				
30	1.00	0.00	1.00	.50	1.00	.50
31-34	1.00	0.00				
35	1.00	0.00			.50	.50
36	1.00	0.00				
37-39	1.00	1.04				
40	1.00	1.04			.50	1.04
41-42	1.00	1.04				
43-44	1.00	1.96				
45	1.00	1.96			1.00	2.96
46-48	1.36	-1.04				
49	1.36	0.00				
50	1.36	0.00			1.36	.50
51-54	1.36	0.00				
55	1.36	1.04			.86	1.54
56-57	1.36	1.04				
58-59	1.80	0.00				
60	1.80	0.00			1.30	0.00

Linking procedure. Lord (1980) described a procedure for removing DIF items prior to linking:

1. Estimate item parameters for all groups combined, standardizing on item difficulty estimates;
2. Reestimate item parameters for each group separately, holding the guessing parameters fixed, and standardizing on item difficulty estimates;
3. Identify DIF items and remove them;
4. Combine groups and estimate θs for examinees;
5. Hold θ fixed and reestimate item difficulty and discrimination for all items for each group separately; and

6. Identify DIF items.

This procedure is difficult because it requires reestimation of item and θ parameters.

Iterative linking of metrics between two groups (Candell & Drasgow, 1988), however, is a somewhat easier procedure to implement:

1. Estimate item parameters independently in each group;
2. Link metrics across groups;
3. Estimate DIF indices for all items on the test and remove items identified as DIF;
4. Relink group metrics using only non-DIF items; and
5. Reestimate DIF indices for all items on the test and remove DIF items.

Iterations over Steps 4 and 5 are continued until either no items are identified as DIF items or the same set of items is identified on a subsequent iteration. The final linking of metrics is based on items that were not identified as DIF.

A major benefit from using this iterative linking procedure is that reestimation of item and θ parameters is not needed. Furthermore, iterative linking has been found to be more accurate than noniterative linking for detecting DIF (Candell & Drasgow, 1988; Kim & Cohen, 1992; Park & Lautenschlager, 1990).

Analysis. Following the linking, all parameter estimates were transformed onto the underlying metric using the TRF method. Two different levels of significance were used to identify DIF—$\alpha = .01$ and $\alpha = .05$. Two different classification errors were possible: false positive errors (FPs) in which a non-DIF item was identified as a DIF item and false negative errors (FNs) in which a DIF item failed to be identified. Parameter recovery was assessed by using root mean square differences (RMSDs) and correlations with the generating parameters.

Results

Recovery of Item and θ Parameters

The recovery results for the first replication are given in Table 3 for both the 20- and 60-item tests for MBE. These results were typical of those obtained across all replications for the different data generation conditions used in this study.

Sample size and test length appeared to affect the magnitudes of both the RMSDs and correlations in each of the generated datasets. Recovery of *b* and θ, however, were less affected by test length than was *a*. Generally, RMSDs for *a* and *b* were smaller and correlations higher for the larger (i.e., $N = 500$) sample than for $N = 100$. RMSDs for θ were larger for the 20-item test than for the 60-item test. Correlations for both *b* and θ were very high in all conditions. RMSDs were larger for MMLE (not shown) for all three parameters.

Based on the results from the recovery analyses, recapture of the underlying item and θ parameters appeared to be very good for the large samples and longer tests and acceptable for the shorter tests in the small samples.

Detection of DIF

Table 4 provides the number of FPs and FNs for each iteration from the iterative linking results from the third replication of MBE parameter estimates for the 60-item test in the 500-examinee sample. For example, for the 20% DIF condition for the 20-Matched group at $\alpha = .01$, iterative linking with *Z(ESA)* required 3 iterations to complete and made 1 FP and 8 FNs. There was very little

Table 3
RMSD and Correlations (r) Between Estimates and True Values of a, b, and θ for 20- and 60-Item Tests for MBE From the First Replication, by Group (N = 100 and 500)

Test Length, Sample Size, and Group	a		b		θ	
	RMSD	r	RMSD	r	RMSD	r
20 Items, N = 100						
Reference	.272	.589	.338	.949	.448	.956
0-Matched	.352	.660	.288	.961	.476	.943
10-Matched	.242	.824	.216	.977	.408	.956
20-Matched	.267	.741	.349	.958	.462	.921
0-Nonmatched	.284	.474	.359	.924	.535	.912
10-Nonmatched	.239	.730	.291	.956	.478	.912
20-Nonmatched	.210	.728	.334	.948	.461	.947
20 Items, N = 500						
Reference	.121	.921	.141	.990	.457	.975
0-Matched	.151	.922	.158	.985	.457	.966
10-Matched	.191	.822	.303	.976	.458	.956
20-Matched	.128	.878	.120	.994	.464	.957
0-Nonmatched	.134	.883	.149	.987	.510	.975
10-Nonmatched	.161	.873	.184	.983	.493	.971
20-Nonmatched	.124	.899	.203	.981	.489	.952
60 Items, N = 100						
Reference	.172	.808	.269	.961	.270	.965
0-Matched	.273	.636	.226	.971	.282	.953
10-Matched	.262	.631	.360	.934	.244	.967
20-Matched	.274	.615	.374	.947	.299	.950
0-Nonmatched	.230	.682	.352	.932	.304	.932
10-Nonmatched	.279	.647	.340	.940	.335	.932
20-Nonmatched	.243	.651	.437	.924	.280	.950
60 Items, N = 500						
Reference	.115	.924	.124	.991	.280	.982
0-Matched	.136	.898	.143	.988	.294	.979
10-Matched	.113	.921	.130	.991	.289	.981
20-Matched	.127	.910	.132	.991	.298	.981
0-Nonmatched	.116	.941	.182	.982	.304	.971
10-Nonmatched	.142	.875	.243	.972	.345	.955
20-Nonmatched	.138	.892	.217	.976	.323	.976

variation in FPs and FNs across replications within each of the simulated conditions.

False Positive Errors

Mean numbers of FPs and FNs calculated from the final iteration results across all five replications are shown in Table 5 for MBE and MMLE. Overall, relatively few differences were observed in the number of FPs made among the three DIF statistics. Lord's χ^2 and *Z(ESA)* tended to have lower mean FPs under most conditions. In addition, FPs appeared to occur slightly more often under the nonmatched-θ DIF conditions for *Z(ESA)* and *Z(H)* than for Lord's χ^2. Mean numbers of FPs also tended to be slightly larger under some conditions with MMLE.

As can be seen in Table 5, the number of possible FPs was clearly a function of test length and the number of non-DIF items in the test. That is, the longer the test, the larger the number of items

Table 4
Numbers of FPs and FNs for Three DIF Statistics for Items on Each Iteration for MBE, From the Third Replication (N = 500 and 60 Items) by Group and α Level

Group, α,	Iteration							
	First		Second		Third		Fourth	
and Statistic	FP	FN	FP	FN	FP	FN	FP	FN
0-Matched, α = .01								
χ^2	0							
ESA	0							
H	0							
0-Matched, α = .05								
χ^2	0							
ESA	0							
H	1		1					
10-Matched, α = .01								
χ^2	0	3	0	2	0	2		
ESA	0	3	0	2	0	2		
H	0	2	0	2				
10-Matched, α = .05								
χ^2	1	2	0	1	0	1		
ESA	2	2	2	2				
H	2	2	2	2	2	2		
20-Matched, α = .01								
χ^2	1	4	0	4	0	4		
ESA	0	8	1	8	1	8		
H	1	6	1	6				
20-Matched, α = .05								
χ^2	4	2	5	1	4	1	4	1
ESA	4	7	2	5	2	5		
H	6	1	3	1	3	1		
0-Nonmatched, α = .01								
χ^2	0							
ESA	0							
H	0							
0-Nonmatched, α = .05								
χ^2	1		1					
ESA	1		1					
H	1		1					
10-Nonmatched, α = .01								
χ^2	1	3	1	3				
ESA	0	4	0	4				
H	0	4	0	4				
10-Nonmatched, α = .05								
χ^2	1	3	1	1	1	1		
ESA	2	4	2	4				
H	2	3	1	2	1	2		
20-Nonmatched, α = .01								
χ^2	1	5	1	5				
ESA	0	9	0	9				
H	0	7	0	7				
20-Nonmatched, α = .05								
χ^2	1	4	1	4				
ESA	3	9	4	8	4	9	4	9
H	4	4	1	4	1	3	1	3

Table 5
Mean Number of FP and FN Identifications From MBE or MMLE for Three DIF Statistics at Two α Levels by Group (N = 100 or 500, With 20 or 60 Items)

Type of Estimation and Group	α = .01						α = .05					
	χ^2		*ESA*		*H*		χ^2		*ESA*		*H*	
	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
MBE: N = 100, 20 Items												
0-Matched	0		.2		0		.4		.8		1.0	
10-Matched	0	1.8	0	2.0	0	2.0	.2	1.6	.8	1.8	.8	1.6
20-Matched	0	4.0	.2	4.0	.2	4.0	.6	3.2	.4	3.6	.4	3.4
0-Nonmatched	0		0		0		0		.2		.6	
10-Nonmatched	0	2.0	0	2.0	.2	2.0	.4	2.0	.6	2.0	.8	2.0
20-Nonmatched	0	4.0	.2	4.0	.2	3.8	.8	3.4	.8	4.0	1.4	3.4
MBE: N = 100, 60 Items												
0-Matched	.4		.2		.2		1.0		.8		1.6	
10-Matched	0	5.4	.2	5.4	.2	5.2	.6	4.8	1.0	4.2	1.0	4.2
20-Matched	0	11.0	0	11.2	0	11.2	.4	9.4	.4	8.6	.8	8.2
0-Nonmatched	0		0		.2		1.0		.8		1.8	
10-Nonmatched	.4	6.0	.8	6.0	.8	6.0	3.0	5.4	3.4	5.2	4.2	5.2
20-Nonmatched	.8	11.0	1.0	11.4	1.2	11.0	1.6	9.6	2.6	11.0	3.2	9.6
MBE: N = 500, 20 Items												
0-Matched	0		0		.4		1.4		.2		1.4	
10-Matched	.4	.6	.2	1.4	.2	1.2	1.0	.2	1.0	.4	1.0	1.0
20-Matched	.2	1.2	0	3.0	.2	2.4	.4	.8	1.2	2.8	1.6	1.8
0-Nonmatched	0		0		0		.6		.6		1.2	
10-Nonmatched	.2	.8	0	2.0	0	1.8	4.0	0	.6	1.0	2.0	1.0
20-Nonmatched	0	1.6	0	3.6	.2	3.0	.6	.6	1.6	3.6	.8	1.4
MBE: N = 500, 60 Items												
0-Matched	0		.2		.2		.4		1.4		3.0	
10-Matched	.2	1.4	.2	1.8	.2	1.6	.6	.8	1.4	.8	3.0	1.2
20-Matched	.2	4.4	.2	8.0	.4	6.6	1.6	1.6	1.2	5.8	3.8	2.0
0-Nonmatched	0		.2		.2		1.8		1.2		2.8	
10-Nonmatched	.6	2.6	.4	4.0	.4	3.4	1.6	2.4	1.6	2.4	2.0	1.2
20-Nonmatched	.6	4.8	.8	9.4	.4	7.6	2.2	3.6	3.2	8.0	3.8	4.0
MMLE: N = 100, 20 Items												
0-Matched	0		0		0		.4		.4		.8	
10-Matched	0	1.8	0	2.0	0	2.0	.4	1.8	.8	1.8	.8	1.6
20-Matched	0	4.0	.2	4.0	.2	3.8	.6	3.2	.4	3.6	.6	3.0
0-Nonmatched	0		0		0		.4		.2		.6	
10-Nonmatched	.4	2.0	.2	2.0	.2	2.0	.4	2.0	1.4	2.0	1.2	2.0
20-Nonmatched	.2	4.0	.4	4.0	.4	3.8	1.4	3.4	1.4	4.0	1.2	3.8
MMLE: N = 100, 60 Items												
0-Matched	.2		0		.2		1.4		.6		1.6	
10-Matched	0	5.6	.2	6.0	.2	6.0	.4	4.6	.4	4.8	1.2	5.0
20-Matched	0	11.2	.2	11.4	.2	11.0	.6	9.2	.4	9.0	.8	8.8
0-Nonmatched	0		0		.4		1.0		.6		1.2	
10-Nonmatched	0	5.8	1.2	6.0	1.0	6.0	3.0	5.4	2.8	5.6	3.6	5.8
20-Nonmatched	.6	11.2	.4	11.4	.8	11.2	2.2	9.2	2.2	11.0	3.4	10.2
MMLE: N = 500, 20 Items												
0-Matched	.2		.2		.4		1.4		.2		1.6	
10-Matched	.4	.8	.2	1.4	.4	1.4	1.2	.6	1.0	.4	1.6	.8
20-Matched	.2	1.6	0	3.2	.2	2.6	.4	.8	1.0	3.0	1.6	2.0

continued on the next page

Table 5, continued
Mean Number of FP and FN Identifications From MBE or MMLE for Three DIF Statistics at Two α Levels by Group (*N* = 100 or 500, With 20 or 60 Items)

Type of Estimation and Group	α = .01						α = .05					
	χ^2		*ESA*		*H*		χ^2		*ESA*		*H*	
	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
0-Nonmatched	.2		0		0		1.4		.4		1.2	
10-Nonmatched	.4	.8	0	2.0	0	1.8	1.0	.4	.8	1.4	1.6	1.2
20-Nonmatched	0	2.2	0	3.6	.2	3.4	.6	.6	1.4	3.6	1.2	2.0
MMLE: *N* = 500, 60 Items												
0-Matched	0		.2		.2		1.0		1.6		2.8	
10-Matched	.2	1.4	.2	2.6	0	1.6	1.0	1.0	1.4	.8	2.6	1.2
20-Matched	.2	4.4	0	8.0	.4	7.2	.2	1.4	1.2	5.8	4.4	3.2
0-Nonmatched	.2		.2		.2		2.2		1.4		3.0	
10-Nonmatched	.6	3.0	.4	4.0	.4	3.4	1.8	2.4	2.4	3.2	3.6	2.8
20-Nonmatched	.6	5.0	1.0	9.6	1.2	8.4	2.4	4.0	3.6	8.8	3.6	5.8

that might be erroneously identified as DIF items. Likewise, the fewer the number of DIF items present on the test, the more items that potentially could be falsely identified as DIF items. Increases in test length, in fact, did result in a slight increase in FPs. This increase was larger for the nonmatched-θ than for the matched-θ conditions.

Another factor influencing the number of FPs was the level of significance. The larger this value, the more items that could potentially fall into the critical region and be falsely identified. This was consistently the case and is illustrated in Figures 1a and 1b for the 60-item test in the 10% DIF conditions from MBE. As the level of significance increased, the mean number of FPs in Table 5 also increased (this increase is also illustrated in Figures 1a and 1b). Finally, the number of FPs in the 0% DIF conditions was lower than the nominal error rate for all three DIF statistics at the α = .01 level, for all conditions for *Z(ESA)*, and for all but one condition for Lord's χ^2 at the α = .05 level.

False Negative Errors

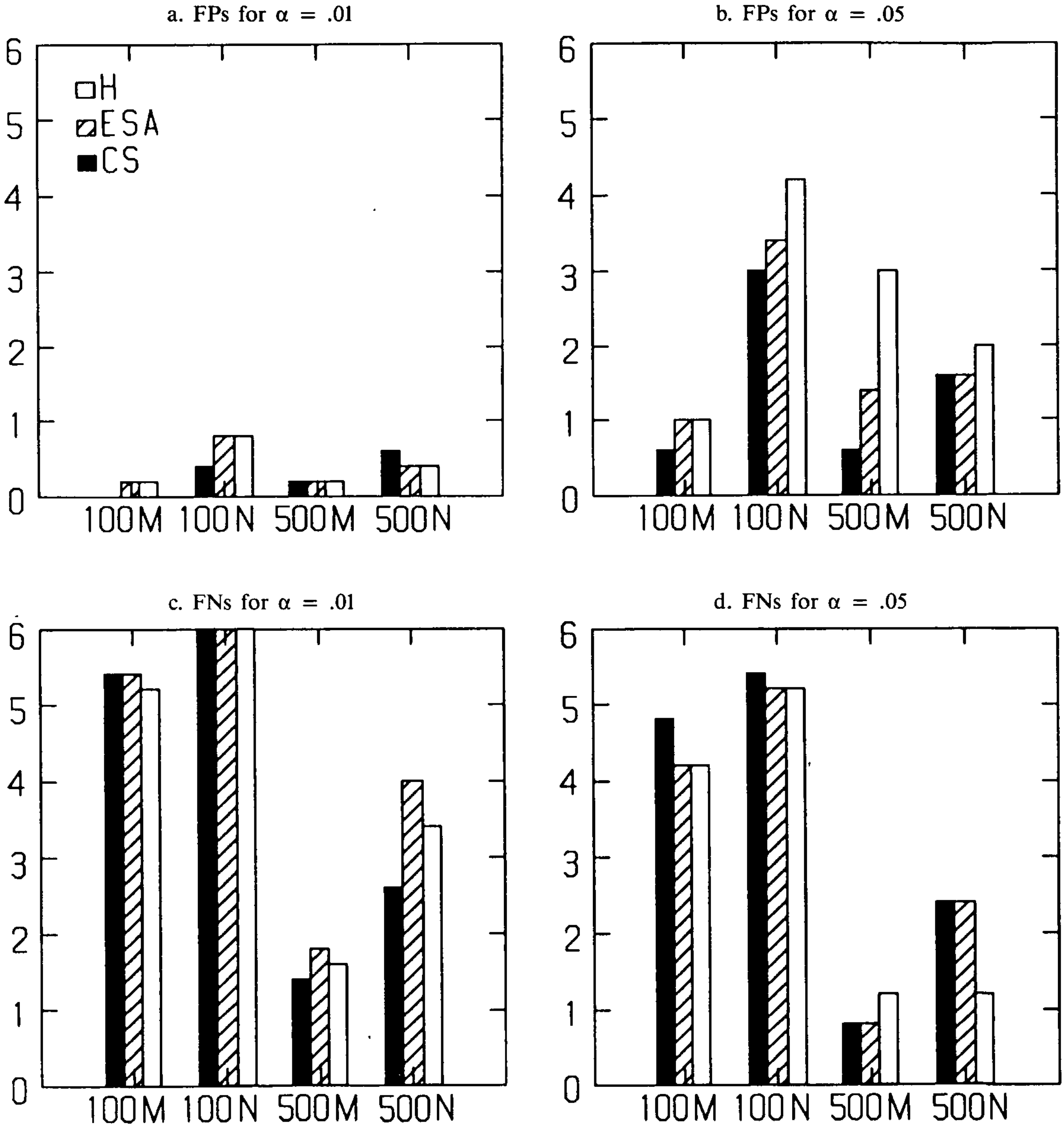
The number of potential FNs is a function, at least in part, of the percent of DIF items on the test; the greater the percentage of DIF items on the test, the greater the possibility of a FN. The present results show that as test length and DIF increased, the mean number of FNs also increased (see Table 5). The mean number of FNs was also larger for the 20% DIF condition, particularly for *Z(ESA)* and *Z(H)*. As test length and sample size increased for each DIF condition, however, the number of FNs decreased.

Overall, fewer FNs were made using either Lord's χ^2 or *Z(ESA)* than using *Z(H)*. The differences between the three statistics, however, were minimal in the *N* = 100 conditions—none of the DIF statistics performed well in that condition. In the *N* = 500 conditions, Lord's χ^2 tended to make the fewest FNs, particularly in the 20% DIF conditions. Differences among the three statistics were largest in the high DIF conditions and favored Lord's χ^2.

Figures 1c and 1d show that for the 60-item test in the 10% DIF, MBE conditions, more FNs occurred in the nonmatched-θ group than in the comparable matched-θ conditions at both levels of significance. Again, however, as is illustrated in Figures 1c and 1d, and as may be seen in Table 5, the differences between the three statistics were not large.

Fewer FNs were found by all three DIF statistics when parameters were estimated with MBE than

Figure 1
FP and FN Rates From MBE for 10% DIF and 60-Item Test with N = 100 and N = 500 for Matched-θ (M) and Nonmatched-θ (N) Groups at α = .01 and α = .05



with MMLE (see Table 5). This difference, although not large, was more evident in the $N = 500$ conditions, particularly for *Z(ESA)* and *Z(H)*. As the recovery results indicate, MBE estimates had smaller RMSDs and higher correlations with the generating parameters. Finally, as would be expected, fewer FNs occurred for each of the three statistics at the .05 level than at the .01 level.

Discussion

Recovery of the underlying item and θ parameters was better for longer tests and for larger samples. Parameter recovery using MBE appeared to be somewhat better than using MMLE in some of the 20-item and 100-examinee conditions.

Relatively few FPs were made using any of the three DIF statistics. The number of FPs increased, as would be expected, as test length and level of significance increased. In the null conditions, the rate of FPs appeared to be within the nominal level of significance. The nonmatched-θ conditions tended to result in more FPs than were observed in the matched-θ conditions. This was more apparent for *Z(ESA)* and *Z(H)* than for Lord's χ^2. The results provide some indication that Lord's χ^2 may result in slightly fewer FPs than either of the other two DIF statistics.

Fewer FNs tended to be made under those conditions of test length, sample size, and parameter estimation algorithm than would be expected to provide more precise item parameter estimates—such as longer tests, larger samples, or use of MBE. Similarly, under conditions that would tend to result in less precise item parameter estimates—such as smaller samples, shorter tests, or high DIF—more FNs tended to be made. FNs tended to be made less frequently using Lord's χ^2 than using *Z(ESA)* or *Z(H)*. As test length and percent of DIF items increased, however, the number of FNs also increased. Likewise, as the sample size increased or the level of significance used for the DIF statistic changed from .01 to .05, the number of FNs decreased. In addition, more FNs were observed under the nonmatched-θ than under the matched-θ conditions. These differences tended to be larger for *Z(ESA)* than for Lord's χ^2. Further, fewer FNs occurred when item parameters were estimated using MBE than when using MMLE.

The relatively strong performance of Lord's χ^2 was interesting, particularly in light of the findings by McLaughlin & Drasgow (1987) that, when both item and θ parameters are unknown, Type I errors may be seriously inflated beyond the nominal level of significance. The apparent rates of errors under the null DIF conditions for both MBE and MMLE in the present study indicated that Type I error control was maintained by Lord's χ^2. One reason for this performance may be due to the impact of marginalized estimation of item parameters; the results from McLaughlin & Drasgow were obtained using joint maximum likelihood estimation. Marginalized solutions provide improved estimates of item parameters over those obtained from joint maximum likelihood estimation (e.g., Drasgow, 1989; Mislevy & Bock, 1986; Mislevy & Stocking, 1990). This improvement in the performance of Lord's χ^2 for both MMLE and MBE in spite of the fact that θ was also treated as unknown is consistent with results reported by Lim & Drasgow (1990).

The effectiveness of Lord's χ^2 over *Z(ESA)* may be due in part to the way the *ESA* is obtained for the 2PM. Raju (1988) defined the *ESA* for the 2PM as $(\hat{b}_F - \hat{b}_R)$. That is, *Z(ESA)* for the 2PM only considers the difference in item difficulties. Lord's χ^2, on the other hand, includes differences between both $(\hat{a}_F - \hat{a}_R)$ and $(\hat{b}_F - \hat{b}_R)$. In addition, the dispersion matrix used for the error term for Lord's χ^2 considers both the variance of the item discrimination as well as the off-diagonal covariance terms; these are not part of the *Z(ESA)* error term.

The results also indicated that *Z(H)* was not as effective as Lord's χ^2. Exactly why this was so is not clear, given that both item discrimination and item difficulty parameter estimates are included in *Z(H)*. It may be that the approximation Raju (1990) used in the computation of *H* does not yield a sufficiently robust statistic for testing DIF under the kinds of conditions simulated in this study.

The generality of results from simulation studies is necessarily compromised to the extent that the generating conditions deviate from real test data. To the degree this does not occur, the results of the study should provide support for the use of Lord's χ^2 over $Z(ESA)$ or $Z(H)$ for detection of DIF when small samples, short tests, nonmatched-θ focal groups, or higher percentages of DIF items are present. When samples are large, tests are long, or the percent of DIF items on a test is not too great (e.g., lower than 10%), then the selection of one of these three DIF statistics may be less important.

References

Baker, F. B. (1986). *GENIRV: A program to generate item response vectors* [Computer program]. Madison WI: University of Wisconsin, Laboratory of Experimental Design.

Baker, F. B., Al-Karni, A., & Al-Dosary, I. M. (1991). EQUATE: A computer program for the test characteristic curve method of IRT equating. *Applied Psychological Measurement, 15,* 78.

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.

Candell, G. L., & Hulin, C. L. (1987). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. *Journal of Cross-Cultural Psychology, 17,* 417–440.

Cohen, A. S., Kim, S. H., & Subkoviak, M. J. (1991). Influence of prior distributions on detection of DIF. *Journal of Educational Measurement, 28,* 49–59.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9,* 413–415.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13,* 77–90.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16,* 209–225.

Kim, S. H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement, 15,* 269–278.

Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29,* 51–66.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5,* 159–173.

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: Erlbaum.

McCauley, C. D., & Mendoza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement, 9,* 389–400.

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.

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1990). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57–75.

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

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197–207.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17,* 1–10.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9,* 93–128.

Shepard, L., Camilli, G., & Williams, D. M. (1985).

Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22,* 77–105.

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

Acknowledgments

The authors express their appreciation to Stacey Ruffalo for assistance in conducting analyses and checking the results of this study and to Lynn Grimm for preparing the figures. The authors also thank the editor and two anonymous reviewers for their many valuable comments on an earlier version of this manuscript. Portions of this paper were presented at the International Symposium on Modern Theories in Measurement: Problems and Issues, University of Ottawa, Ontario, Canada, November, 1991.

Author's Address

Send requests for reprints or further information to Allan S. Cohen or Seock-Ho Kim, University of Wisconsin, 1025 W. Johnson, Madison WI 53706, U.S.A.