An Investigation of the Likelihood Ratio Test For Detection of Differential Item Functioning

Allan S. Cohen, University of Wisconsin, Madison

Seock-Ho Kim, The University of Georgia

James A. Wollack, University of Wisconsin, Madison

Type I error rates for the likelihood ratio test for detecting differential item functioning (DIF) were investigated using monte carlo simulations. Two- and three-parameter item response theory (IRT) models were used to generate 100 datasets of a 50-item test for samples of 250 and 1,000 simulated examinees for each IRT model. Item parameters were estimated by marginal maximum likelihood for three IRT models: the three-parameter model, the three-parameter model with a fixed guessing parameter, and the two-parameter model. All DIF comparisons were simulated by randomly pairing two samples from each sample size and IRT model condition so that, for each sample size and IRT model condition, there were 50 pairs of reference and focal groups. Type I error rates for the two-parameter model were within theoretically expected values at each of the α levels considered. Type I error rates for the three-parameter and three-parameter model with a fixed guessing parameter, however, were different from the theoretically expected values at the α levels considered. *Index terms: bias, differential item functioning, item bias, item response theory, likelihood ratio test for DIF.*

An item is said to be functioning differentially when the probability of a correct response is different for examinees at the same trait level but from different groups (Pine, 1977). Because the presence of such items on a test is a threat to validity and may seriously interfere with efforts to equate tests, they must be removed from consideration. Thissen, Steinberg, & Gerrard (1986) and Thissen, Steinberg, & Wainer (1988, 1993) proposed the likelihood ratio test (LR; Neyman & Pearson, 1928) to evaluate the significance of observed differences in item responses from different groups under item response theory (IRT). However, little evidence has been presented regarding the effectiveness of LR in this context. Thissen et al. (1988, 1993) provided an excellent discussion of the theoretical relationship between LR and Lord's (1977, 1980) χ^2 test (LC) for differential item functioning (DIF) but included only a few illustrative examples. Kim & Cohen (1995) also have presented some data that point to the effectiveness of LR for DIF detection. No information has yet been presented, however, regarding Type I error rates for LR in DIF detection. The present study was designed to examine Type I error rates of LR for DIF detection (Thissen et al.) under three IRT models for dichotomously scored items.

The basic building block of IRT is the item response function (IRF). The IRF of IRT models for dichotomously scored items describes the functional relationship between the probability of a correct response to an item and examinee trait level (θ). In the context of IRT, an item functions differentially if the IRFs obtained from different groups of examinees are different (Lord, 1980). IRFs can be identical if and only if the sets of item parameters estimated in different groups are equal.

The equality of item parameters can be tested using several different approaches under IRT. One approach is to compare item parameters estimated from two groups of examinees (e.g., Draba, 1977; Lord, 1977, 1980; Wright & Stone, 1979). A second approach is to compare IRFs estimated from two

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 1, March 1996, pp. 15–26
© Copyright 1996 Applied Psychological Measurement Inc.
0146-6216/96/010015-12$1.85

groups of examinees by measuring the area between them (e.g., Kim & Cohen, 1991; Linn, Levine, Hastings, & Wardrop, 1981; Raju, 1988, 1990; Rudner, 1977; Wainer, 1993). A third approach is to compare likelihood functions, using a likelihood ratio, to evaluate the differences between item responses from two groups (e.g., Thissen et al., 1986, 1988, 1993). Thissen et al. (1988) noted that this LR approach is preferable for theoretical reasons because the first two approaches may require accurate estimates of variances (and covariances) of the item parameters. At the present time, computational difficulties continue to impede obtaining accurate estimates of these variances (and covariances).

DIF studies under IRT require that estimates of item parameters obtained in different groups first be placed on a common metric before comparisons are made (Haebara, 1980; Kim & Cohen, 1992; Stocking & Lord, 1983). DIF detection based on statistics such as LC or Raju's (1988, 1990) area measures accomplish this by first calibrating item parameters in different groups and then subsequently using some method for transforming the parameter estimates onto a common metric. For LR, using the computer program MULTILOG (Thissen, 1991) makes such transformations unnecessary because item parameters are estimated simultaneously in each group.

For LR, the problem of a common metric is handled through the common or anchor set of items used rather than by equating (Millsap & Everson, 1993). In LR (Thissen et al., 1988, 1993), the likelihood from a compact model, in which no group differences are assumed to be present, is compared to that from an augmented model, in which one or more items are examined for possible DIF. The metrics of the compact and augmented models are dependent on the anchor items. The tentative assumption is that there are no DIF items among the common items. That is, the anchor set for each augmented model is assumed to have no items that function differentially in either group. Note that comparing a compact model to an augmented model requires two separate calibrations to obtain the likelihoods for each comparison—one for the compact model and one for the augmented model.

The presence of DIF items among the common items used to link metrics may seriously interfere with DIF detection (Kim & Cohen, 1992; Shepard, Camilli, & Williams, 1984). Methods for removing such items are available for all three DIF detection approaches. Lord (1980) proposed a single iteration method for purification of the linking items with a χ^2 test comparing item parameters. A simpler technique called iterative linking (Candell & Drasgow, 1988) has been used to remove DIF items from the linking items with DIF detection methods that compare item parameters estimated in both groups (Cohen & Kim, 1993; Kim & Cohen, 1992; Park & Lautenschlager, 1990) and with methods that measure areas between IRFs (Cohen & Kim, 1993). For LR, Thissen et al. (1988, 1993) recommended using the Mantel-Haenszel χ^2 (Holland & Thayer, 1988) to remove potential DIF items from the common items. Kim & Cohen (1995) recommended using an iterative purification method for detection and removal of DIF items from the anchor set. The Kim and Cohen method uses an iterative elimination process in which LR is used to detect DIF in each item of the test. After each iteration, the item with the largest significant χ^2 is removed and the next iteration begun. Iterations continue until no additional DIF items are detected. Kim and Cohen reported results from this method that were very close to those obtained using iterative linking with item parameter comparison and area measures. In the present study, only the number of significant LR χ^2s obtained from the first stage of this iterative purification process was investigated because no DIF was simulated in these data.

Method

Data Generation

Two sample sizes were used in order to simulate a small sample ($N = 250$) and a large sample ($N = 1{,}000$) condition. Both the three-parameter logistic model (3PLM) and the two-parameter logistic model (2PLM) were used to generate the datasets. The probability of a correct response for an examinee on item i ($i = 1, ..., n$) for the 3PLM is given by

$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp[1.7a_i(\theta - b_i)]}{1 + \exp[1.7a_i(\theta - b_i)]}, \qquad (1)$$

where
b_i is the item difficulty parameter,
a_i is the item discrimination parameter,
c_i is the pseudoguessing parameter,
θ is the trait level parameter, and
1.7 is a scaling factor used to transform the metric from logistic to normal.

The probability of a correct response for an examinee on item i for the 2PLM is given by

$$P_i(\theta) = \frac{\exp\left[1.7a_i(\theta - b_i)\right]}{1 + \exp\left[1.7a_i(\theta - b_i)\right]} . \qquad (2)$$

The datasets used in this study were the same as those used in Kim, Cohen, & Kim (1994). For each of the four combinations of sample size (250 or 1,000) × IRT model (2PLM or 3PLM), 100 datasets of a 50-item test were simulated using the computer program GENIRV (Baker, 1988). Thus, 400 datasets were generated.

θs were sampled from the standard normal distribution, $\theta_j \sim N(0,1)$. The item numbers and generating parameters for the 3PLM are shown in Table 1. For the 2PLM, only the a and b parameters were used. These values, originally reported by Lord (1968), were used by both McLaughlin & Drasgow (1987) and Kim et al. (1994) in two previous studies of Type I error rates for LC, thus allowing comparisons with the results from those studies. Such comparisons were of interest because both LR and LC are asymptotically equivalent (Thissen et al., 1988).

In a typical DIF study, there are two groups of examinees—the focal group and the reference group. The reference group is considered the base group against which the focal group is compared. In the present study, DIF comparisons were simulated by randomly pairing two sets of data from the same IRT model and sample size conditions and identifying one group as the reference group and the other as the focal group. Of the 400 datasets generated, 200 pairs of reference and focal groups were obtained (i.e., 50 pairs for each sample size and IRT model combination). No DIF comparisons were simulated between different size samples.

Item Parameter Estimation

Each pair of reference and focal groups was analyzed using the default options available in the marginal maximum likelihood estimation (MMLE) algorithm implemented in the computer program MULTILOG (Thissen, 1991). The 2PLM was used to estimate item parameters in the datasets generated using the 2PLM. The 3PLM and the 3PLM with fixed c (3PLM-c) were used to estimate item parameters in the datasets generated with the 3PLM. The 3PLM-c model described by Thissen et al. (1988, 1993) constrains the c for each item to be equal in the reference and focal groups. In Kim et al. (1994), however, c was fixed for all items to .15, the average of the c across all 50 items. The Thissen et al. approach was used here.

The Likelihood Ratio Test

LR (Thissen et al., 1988, 1993) compares two different models—a compact model and an augmented model. LR is the difference between the values of −2 times the log likelihood for the compact model (L_C) and −2 times the log likelihood for the augmented model (L_A). The likelihood, L, can be obtained from the output of MULTILOG and is based on the results over the entire dataset following MMLE estimation. LR (called G^2 by Thissen et al.) can be written as

$$\begin{aligned}\mathrm{LR} &= -2\log L_C - (-2\log L_A) \\ &= -2\log L_C + 2\log L_A . \end{aligned} \qquad (3)$$

LR is distributed as a χ^2 under the null hypothesis with degrees of freedom (df) equal to the difference in the number of parameters estimated in the compact and augmented models. For this study, LR was distributed as a χ^2 with 2 df for DIF comparisons under the 2PLM and 3PLM-c and 3 df for DIF comparisons under the 3PLM.

In the compact model, the item parameters are assumed to be the same for both the reference and focal groups. MULTILOG has an option that permits equality constraints to be placed on items for estimation of the compact model. In this study, the parameter estimates for all 50 items in the compact

Table 1
Generating Item Parameters (a, b, and c) and Number of Significant LRs at $\alpha = .05$ for Three Models (2PLM, 3PLM-c, and 3PLM) and Two Sample Sizes ($N = 250$ and 1,000)

				2PLM		3PLM-c		3PLM	
Item	a	b	c	250	1,000	250	1,000	250	1,000
1	1.1	−.7	.20	4	6	23	4	3	4
2	.7	−.6	.20	1	5	9	3	4	3
3	1.4	.1	.20	1	1	0	2	4	3
4	.9	.9	.16	1	4	6	4	2	7
5	1.2	.7	.12	2	8	3	7	4	3
6	1.6	1.1	.06	2	2	9	4	0	6
7	1.6	1.1	.06	5	5	2	6	4	2
8	1.6	−.1	.16	2	5	3	4	4	5
9	1.2	.5	.20	2	1	3	5	1	3
10	2.0	1.6	.16	5	3	4	6	5	4
11	1.0	1.6	.13	2	1	3	1	3	3
12	1.5	1.7	.09	2	1	5	1	3	4
13	1.0	.7	.15	3	4	5	3	3	4
14	1.1	2.0	.06	1	5	3	3	4	4
15	1.1	2.4	.09	6	1	4	1	1	2
16	2.0	1.4	.11	4	2	6	3	3	6
17	1.7	1.3	.17	0	1	1	4	5	1
18	.5	−.6	.20	5	4	4	6	6	1
19	.9	1.6	.11	1	1	3	8	5	4
20	1.3	.4	.18	3	2	4	1	1	4
21	1.1	1.2	.05	4	1	7	4	4	2
22	1.2	1.1	.05	2	4	2	3	6	2
23	1.3	.2	.20	2	5	5	1	6	4
24	1.3	.2	.20	5	5	4	3	3	6
25	.5	−.8	.20	4	3	2	4	5	4
26	.7	.5	.20	4	5	5	3	1	2
27	.7	.5	.20	2	1	5	1	4	4
28	.4	−.4	.20	0	3	5	2	3	2
29	.4	−.4	.20	2	0	1	2	2	2
30	1.2	−.5	.20	2	3	2	10	5	8
31	.7	−1.0	.20	0	2	2	2	5	2
32	.7	−.2	.20	2	5	3	4	4	1
33	.7	−.2	.20	1	0	1	7	2	2
34	.5	0.0	.20	4	3	8	6	4	5
35	.9	.5	.14	4	5	6	4	3	0
36	1.1	1.4	.04	2	1	3	9	1	3
37	1.2	−.6	.20	1	0	1	3	1	2
38	1.2	−.6	.20	1	5	5	2	1	3
39	.6	−.5	.20	2	1	6	5	5	4
40	1.6	.3	.18	2	2	4	4	6	1
41	1.1	0.0	.20	3	1	5	4	4	3
42	1.5	2.0	.06	3	3	9	5	4	7
43	1.9	1.9	.11	4	3	3	1	2	4
44	.9	−.5	.20	7	2	0	4	4	1
45	.7	−.5	.20	2	3	3	1	2	2
46	1.4	1.6	.11	4	3	1	3	4	2
47	1.4	1.6	.11	1	1	4	4	5	4
48	1.0	1.7	.08	2	2	4	4	3	6
49	1.2	1.1	.15	4	3	1	2	4	1
50	1.2	1.1	.15	3	2	4	5	7	2
Total				131	139	211	188	175	164

model were set to be equal in both the reference and focal groups. In the augmented model, item parameters for all items except the studied item(s) were constrained to be equal in both the reference and focal groups. These constrained items are referred to as the common or anchor set. In a DIF comparison, in other words, only the item parameters for the studied item(s) are estimated separately in the reference and focal groups. For example, in this study, for the augmented model in which Item 1 was the studied item, item parameter estimates for Item 1 were unconstrained in the reference and focal groups. Items 2–50 formed the anchor set for this augmented model and so were each constrained to have the same parameter estimates in both groups. The metric used in LR, therefore, is based on the set of items contained in the anchor set. In this study, the augmented models were constructed to study a single item at a time. All items were studied sequentially for DIF.

Error Rates

Error rates were obtained by comparing the number of significant LRs to the total number of augmented model calibration runs conducted for a given IRT model and sample size condition. For a single test, 51 separate calibration runs were required to estimate the necessary likelihood statistics—one run to estimate the likelihood for the compact model and 50 runs for each of the augmented models (i.e., one augmented model for each of the 50 items). For the 50 pairs of reference and focal groups in a sample size × IRT model condition, 2,550 separate model calibration runs were required. For the three IRT models × sample size conditions, a total of 15,300 calibrations were required.

Results

Table 1 shows the number of significant LRs for each item at $\alpha = .05$. The data in this table illustrate the general pattern of results obtained. A similar pattern of results was found at all other α levels examined.

For Item 1, estimation with the 2PLM resulted in 4 significant LRs for $N = 250$ and 6 for $N = 1{,}000$. For this same model, there were a total of 131 significant LRs across all 50 items for $N = 250$ and 139 for $N = 1{,}000$. Table 2 shows that the Type I error rate for the 2PLM was .0524 for $N = 250$ and .0556 for $N = 1{,}000$. The expected number of significant LRs for a single item over 50 replications at $\alpha = .05$ would be 2.5. For 50 items, the expected number of significant LRs over 50 replications would be 125, assuming all item parameter estimations converged successfully.

Table 2
Proportion of Significant LRs for the 2PLM, 3PLM-*c*, and 3PLM Across all Samples and All Items at α Levels From .005 to .10

Model and α	$N = 250$	$N = 1{,}000$
2PLM		
.0005	.0008	.0004
.001	.0012	.0012
.005	.0060	.0072
.01	.0100	.0104
.05	.0524	.0556
.10	.0912	.1084
3PLM-*c*		
.0005	.0101	.0216
.001	.0141	.0224
.005	.0262	.0268
.01	.0330	.0325
.05	.0849	.0753
.10	.1308	.1178
3PLM		
.0005	.0245	.0048
.001	.0253	.0081
.005	.0277	.0141
.01	.0329	.0217
.05	.0701	.0660
.10	.1246	.1159

The default option for MMLE in MULTILOG empirically selects starting values to begin the calibration. Failure to converge using these starting values occurred in less than 10% of the 2,550 calibration runs for any sample size × IRT model combination. When MULTILOG failed to converge for any of the compact or augmented models, the generating parameters given in Table 1 were used as starting values for that item in that particular dataset. When the generating values were used as starting values, item parameter estimates were obtained for nearly all previously nonconverged items: All item parameter estimation runs converged successfully for the 2PLM; for the 3PLM-*c*, 15 cases for $N = 250$ and 4 cases for $N = 1{,}000$ failed to converge; for the 3PLM, 5 cases for $N = 250$ and 16 cases for $N = 1{,}000$ failed to converge. The results presented in Tables 1 and 2 are for items that converged using either the default starting values determined by MULTILOG or using generating values as starting values. Items for the 3PLM-*c* and 3PLM that did not converge were excluded from subsequent calculations.

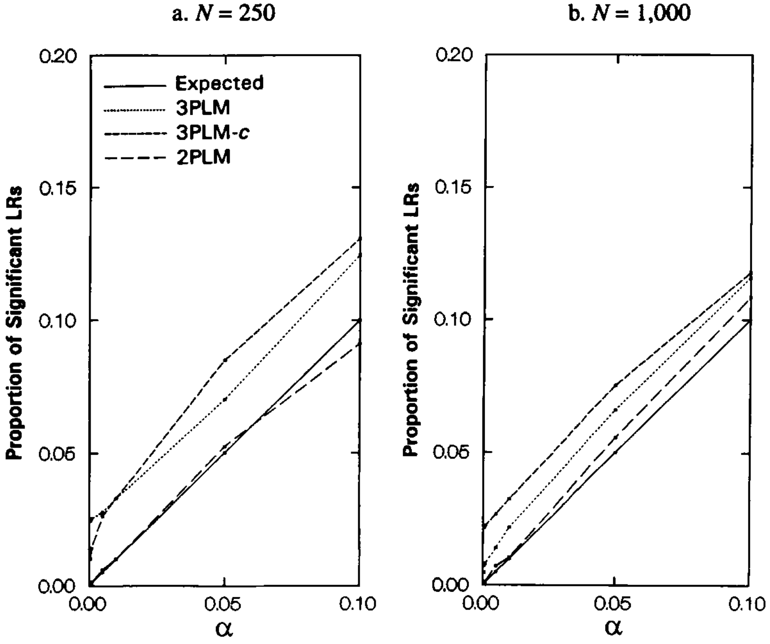
Type I Error Rates

The Type I error rates presented in Table 2 are the percentages of significant LRs over all replications. Differences in error rates for the 2PLM, 3PLM-*c*, and 3PLM are illustrated in Figures 1a and 1b, for $N = 250$ and 1,000, respectively.

There were consistent differences in Type I error rates depending on the IRT model. Figures 1a and 1b show that Type I error rates for the 2PLM were always close to the nominal α levels for both sample sizes. For the 3PLM-*c* and 3PLM, however, error rates were inflated for all α levels. For the 3PLM-*c*, at the .0005 level, the error rate was .0101 which is approximately 20 times higher than expected for $N = 250$, and .0216 which is 43 times higher for $N = 1{,}000$; at the .001 α level, the error rate was .0141 which is approximately 14 times higher than expected for $N = 250$, and .0224 which is 22 times higher for $N = 1{,}000$. Similar results were obtained at these α levels for the 3PLM and $N = 250$. For $N = 1{,}000$, however, error rates for these same α levels for the 3PLM were much lower, albeit still inflated over nominal α levels. At the .005 α level, the error rate for the 3PLM-*c* was approximately 5 times the nominal rate for both $N = 250$ (i.e., the error rate was .0262) and $N = 1{,}000$ (i.e., the error rate was .0268). The difference between error rates and theoretically expected values narrowed somewhat at α levels from .01 to .10. A similar pattern was observed for the 3PLM for both $N = 250$ and 1,000.

Sample size did not appear to have a consistent effect on error rates. The error rates for the 2PLM (Figure 2a) were close to the expected rate at each of the

Figure 1
Proportion of Significant LRs for Two Sample Sizes



α levels examined for both sample sizes. For the 3PLM-*c* (Figure 2b), at nominal α levels of .0005, .001, and .005, error rates were slightly smaller for $N = 250$ than those for the 3PLM (Figure 2c). At all other α levels in both sample size conditions, the error rates for the 3PLM were slightly closer to the expected levels than those for the 3PLM-*c*. For example, for $\alpha = .0005$ and $N = 250$, the error rate for the 3PLM-*c* was .0101 compared to .0245 for the 3PLM. At $\alpha = .10$, however, the error rate of the 3PLM-*c* was .1308 compared to .1246 for the 3PLM.

Relationships Among Generating Parameters and Significant LRs

Table 3 shows the Spearman rank-order correlations (ρ) between the generating parameters and the number of significant LRs. (Correlations with the *c* parameter were applicable only to the 3PLM.)

For all three models, none of the ρs between the number of significant LRs and item parameters were significant for either sample size. These results are in marked contrast to previous results in which a substantial number of significant ρs were observed between the generating parameters and the number of significant LCs (Kim et al., 1994; McLaughlin & Drasgow, 1987). In the present study, the number of significant LRs showed no consistent relationship to the values of the generating item parameters.

Comparison of Type I Errors

Type I error rates at the .05 α level were compared with results from Kim et al. (1994). ρs between significant LCs for marginal Bayesian estimation (MBE) and MMLE results (from Kim et al.) and significant LRs obtained in this study are reported in Table 4.

None of the ρs between the number of significant LCs and LRs for the 3PLM, which ranged from 0.0 to −.181 for $N = 250$ and from 0.0 to −.028 for $N = 1{,}000$, were significantly different from 0.0. A number of significant ρs were observed, however, between LC and LR for the 3PLM-*c* and 2PLM. For the 3PLM-*c* and 2PLM, ρs for $N = 250$ and $N = 1{,}000$ were significant (e.g., for the 2PLM and $N = 250$, the ρ between LR and LC was .402 for MBE and .424 for MMLE). These results are in general agreement

Figure 2
Proportion of Significant LRs for Three Models

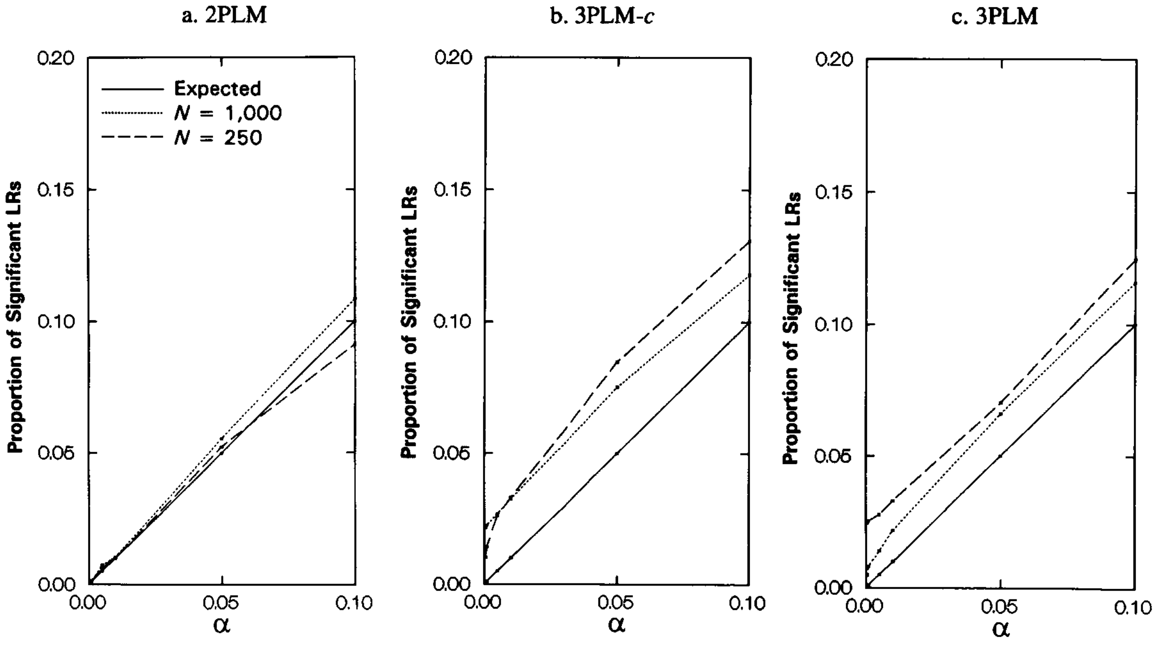


Table 3
Spearman ρs Among Generating Item Parameters and Number of Significant LRs at α = .05 for the 3PLM, 3PLM-*c*, and 2PLM

	Generating Parameter		
	a	*b*	*c*
Generating Parameters			
b	.500**		
c	–.474**	–.847**	
3PLM			
$N = 250$	.235	.202	–.142
$N = 1{,}000$	.046	–.064	.034
3PLM-*c*			
$N = 250$	–.064	.064	
$N = 1{,}000$	–.019	–.025	
2PLM			
$N = 250$	.113	.143	
$N = 1{,}000$	.025	–.166	

**$p < .01$.

with those reported by Kim & Cohen (1995), which indicated strong similarities between these two DIF detection measures for the 2PLM. Correlations between sample sizes, however, tended not to be significantly different from 0.0 (except for the .431 for the 2PLM).

Type I error rates are also plotted in Figures 3a–3f for the 3PLM, 3PLM-*c*, and 2PLM for both sample sizes, along with results from Kim et al. (1994) for MMLE and MBE. The solid line in each figure represents the expected number of significant LRs and LCs at various α levels.

Both McLaughlin & Drasgow (1987) and Kim et al. (1994) reported serious loss of Type I error control with the 3PLM for LC. This problem is clearly evident in Figures 3a and 3b for both MBE and MMLE. LR results from the present study present a markedly different picture: Type I errors for the 3PLM were much closer to theoretically expected values for both $N = 250$ and $N = 1{,}000$.

Results for the 3PLM-*c* (Figures 3c and 3d) for LC under MBE and MMLE were closer to theoretically expected values at α = .05 than for the 3PLM. Type I errors for LC were lower than expected, whereas errors for LR were inflated. Recall that in this study the *c*s were constrained to be equal for each item separately, whereas the constraint used by McLaughlin & Drasgow (1987) and Kim et al. (1994) for LC maintained the same value of *c* for all items on the test.

Results for the 2PLM and $N = 250$ (Figure 3e) indicate that error rates for LC were below expected values. Results for the 2PLM and $N = 1{,}000$ (Figure 3f) indicate that error rates for LC were slightly below, albeit quite close to, expected values. Results for LR for both sample sizes, however, were essentially at (or slightly above) expected values.

Discussion

Type I error rates for LR for the 2PLM were very close to those expected for both sample size conditions at each of the α levels considered. For the 3PLM and 3PLM-*c*, error rates at the .0005 to .005 levels were higher than nominal α levels but error rates at the .01 to .10 levels were closer to expectation.

Comparisons of error rates with those of LC at α = .05, derived from previous research, indicated good agreement within sample size for the 2PLM and 3PLM-*c* models. Lack of agreement for the 3PLM was observed; LC results for this model were markedly divergent from theoretical expectations whereas those for LR were somewhat less so. The results suggest that for all three models, Type I error rates in general for LR were closer to theoretically expected values than for LC.

The primary concern in most DIF studies is to be able to detect all items that function differentially. This is normally accomplished by setting the α level high—for example, at .05, .10, or even higher. At such α levels, LR provided excellent Type I error control for the 2PLM and reasonable control for the 3PLM-*c* and 3PLM. In DIF studies, however, there is also a concern for the power of the DIF statistic; that is, the extent to which it provides control over Type II errors. Such errors occur when DIF items fail to be detected. Cohen & Kim (1993) demonstrated that iterative linking (Candell & Drasgow, 1988) used with LC provided good Type II error control with the 2PLM. Power studies of LR for DIF are also needed.

In this study, the underlying θ distributions for the reference and focal groups were identical and, consequently, the LRs were performed under relatively ideal conditions. Previous research (Cohen & Kim, 1993) has indicated that more Type I errors were observed for LC and Raju's area measures

Table 4
Spearman ρs Between Number of Significant LRs at $\alpha = .05$ and Number of Significant LCs at $\alpha = .05$, for $N = 250$ and $N = 1,000$

Model, Statistic, *N*, and Estimation	LC				LR
	N = 250		*N* = 1,000		
	MBE	MMLE	MBE	MMLE	*N* = 250
3PLM					
LC					
250					
MBE					
MMLE	.461**				
1,000					
MBE	.464**	.730**			
MMLE	.362**	.831**	.837**		
LR					
250	0.000	−.091	−.048	−.181	
1,000	−.028	−.021	0.000	−.003	−.101
3PLM-*c*					
LC					
250					
MBE					
MMLE	.582**				
1,000					
MBE	.024	−.018			
MMLE	.043	.002	.897**		
LR					
250	.289*	.409**	−.004	.008	
1,000	.036	.009	.546**	.556**	.009
2PLM					
LC					
250					
MBE					
MMLE	.891**				
1,000					
MBE	.333*	.480**			
MMLE	.335*	.489**	.996**		
LR					
250	.402**	.424**	.039	.057	
1,000	.262	.431**	.788**	.790**	.240

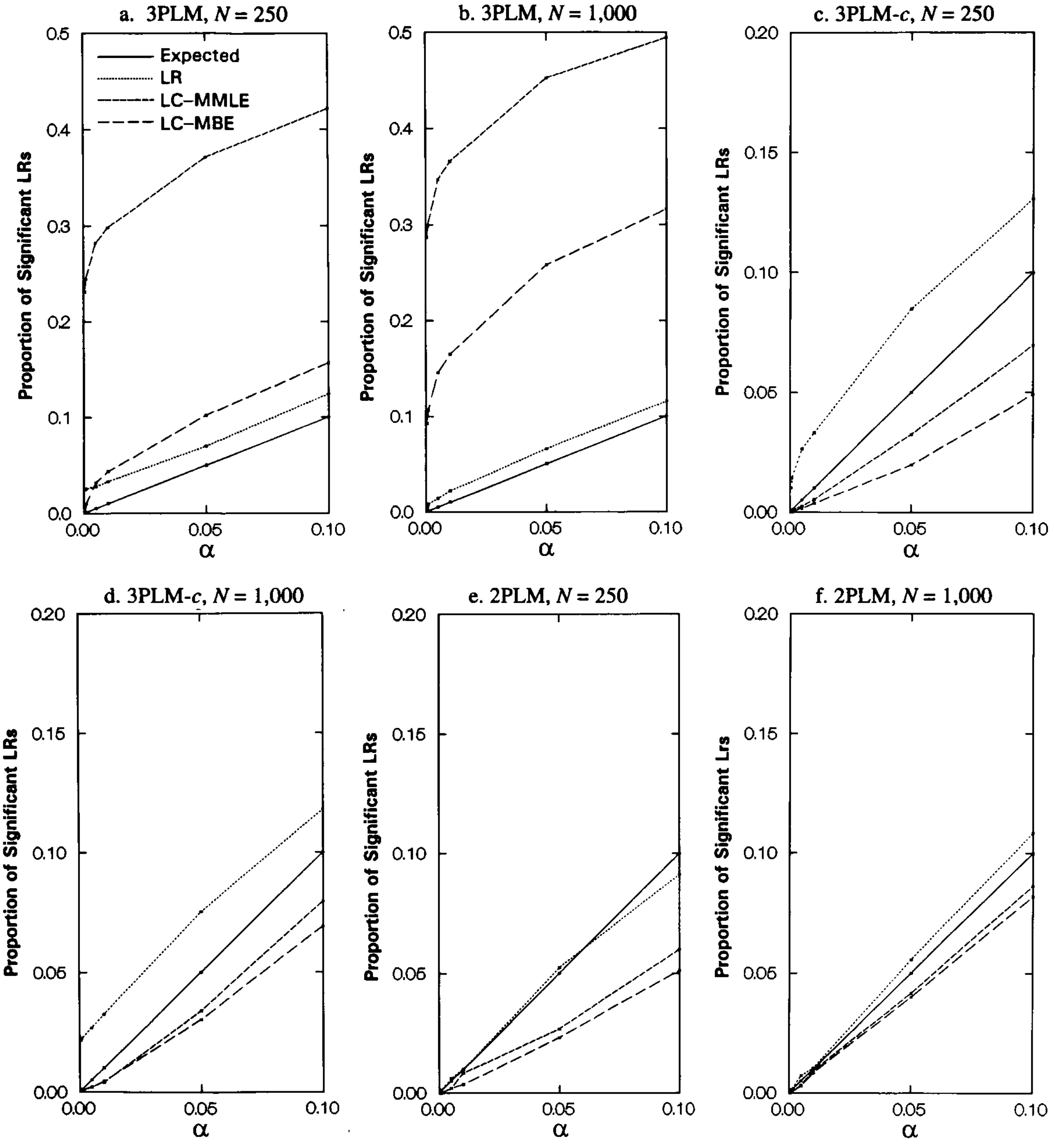
**p* < .05; ***p* < .01.

when the two θ distributions were not matched. It would seem reasonable to suspect that differences in the underlying θ distributions might induce errors in item parameter estimation and subsequent LRs as well. This issue was not addressed in the present study.

Errors in DIF detection due to the presence of DIF items in the linking set have been reported with parameter comparison and area DIF measures (Kim & Cohen, 1992). The effect of the presence of DIF in the anchor items in the augmented model, however, has not yet been studied for LR. Thissen et al. (1988, 1993) recommended using the Mantel-Haenszel χ^2 (Holland & Thayer, 1988) to purify the anchor set prior to calculation of LR between the compact and augmented model. More recently, Kim & Cohen (1995) have described an iterative purification procedure for LR. Further studies with such procedures are important.

LR is an asymptotic statistic and, consequently, it

Figure 3
Proportion of Significant LRs for Three Models and Two Sample Sizes



a. 3PLM, $N = 250$

b. 3PLM, $N = 1{,}000$

c. 3PLM-c, $N = 250$

d. 3PLM-c, $N = 1{,}000$

e. 2PLM, $N = 250$

f. 2PLM, $N = 1{,}000$

would be expected that large sample sizes would be required to obtain satisfactory results. Results from the present study, however, did not indicate consistent differences in error rates due to sample size. Error rates at α levels from .01 to .10 were quite close to expected values for both sample sizes for the 2PLM and reasonably close at $\alpha = .05$ and $\alpha = .10$ for the 3PLM-*c* and 3PLM. If subsequent research finds LR to have adequate power for DIF detection, the major deterrent to its use would appear to be that it requires extensive manipulation of the data. For example, in the present study, for one set of data for the reference and focal groups for a 50-item test, 51 separate MULTILOG calibration runs and an additional specialized routine were required to obtain each of the 50 LRs (i.e., one run for the compact model and one run for each of the 50 augmented models). The same set of data, however, required only two calibration runs using either BILOG (Mislevy & Bock, 1990) or MULTILOG to obtain the reference and focal group item parameter estimates, and a single program to equate item parameters to a common metric and then calculate LC. Software implementing LR would likely make this test a more valuable tool for DIF detection.

References

Baker, F. B. (1988). *GENIRV: Computer program for generating item responses* [Computer program]. Madison: University of Wisconsin, Department of Educational Psychology, Laboratory of Experimental Design.

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.

Cohen, A. S., & Kim, S.-H. (1993). A comparison of Lord's χ^2 and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17,* 39–52.

Draba, R. E. (1977). *The identification and interpretation of item bias* (Research Memorandum No. 25). Chicago: The University of Chicago, Department of Education, Educational Statistics Laboratory.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144–149.

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

Kim, S.-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,* 551–566.

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

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.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989–1020.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam, The Netherlands: Swets & Zeitlinger.

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

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

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

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I and Part II. *Biometrika, 20A,* 174–240, 263–294.

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.

Pine, S. M. (1977). Applications of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing: Proceedings of a symposium presented at the 18th annual convention of the Military Testing Association* (Research Rep. No. 77-1, pp. 37–43). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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 Measure-*

ment, 14, 197–207.

Rudner, L. M. (1977, April). *An approach to biased item identification using latent trait measurement theory.* Paper presented at the annual meeting of the American Educational Research Association, New York.

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

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 207–210.

Thissen, D. (1991). *MULTILOG user's guide* [Computer program]. Chicago: Scientific Software.

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 differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale NJ: Erlbaum.

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

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

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Acknowledgments

Portions of this paper were presented at the annual meeting of the National Council on Measurement in Education at San Francisco, April, 1995.

Authors' Addresses

Send requests for reprints or further information to Allan S. Cohen or James A. Wollack, University of Wisconsin, 1025 W. Johnson, Madison WI 53706, U.S.A. or to Seock-Ho Kim, University of Georgia, 325 Aderhold Hall, Athens GA 30602, U.S.A.