Lord's Chi-Square Test of Item Bias With Estimated and With Known Person Parameters

Mary E. McLaughlin and Fritz Drasgow
University of Illinois

Properties of Lord's chi-square test of item bias were studied in a computer simulation. θ parameters were drawn from a standard normal distribution and responses to a 50-item test were generated using SAT-V item parameters estimated by Lord. One hundred independent samples were generated under each of the four combinations of two sample sizes (N = 1,000 and N = 250) and two logistic models (two- and three-parameter). LOGIST was used to estimate item and person parameters simultaneously. For each of the 50 items, 50 independent chi-square tests of the equality of item parameters were calculated. Proportions of significant chi-squares were calculated over items and samples, at alpha levels of .0005, .001, .005, .01, .05, and .10. The overall proportions significant were as high as 11 times the nominal alpha level. The proportion significant for some items was as high as .32 when the nominal alpha level was .05. When person parameters were held fixed at their true values and only item parameters were estimated, the actual rejection rates were close to the nominal rates.

A major problem in practical uses of item response theory (IRT) is the estimation of item and person parameters. Frequently, parameters are estimated by the method of maximum likelihood. However, when both item and person parameters are estimated simultaneously, the estimates do not have the usual statistical properties of consistency and asymptotic efficiency. The adequacy of estimates can nonetheless be assessed by their performance in practical applications. In this paper the effects of simultaneous estimation on the properties of Lord's (1980, pp. 217–223) chi-square test of item bias are assessed.

Lord's test of item bias may be used with data that fit either two- or three-parameter logistic item response models. Lord's chi-square tests the hypothesis that the item parameters in one subpopulation are equal to those in a second subpopulation. The matrix formulation for the ith item is

$$\chi_i^2 = \mathbf{v}_i' \boldsymbol{\Gamma}_i^{-1} \mathbf{v}_i \quad , \tag{1}$$

where

$$\mathbf{v}_i' = [\hat{b}_{i\mathrm{A}} - \hat{b}_{i\mathrm{B}},\ \hat{a}_{i\mathrm{A}} - \hat{a}_{i\mathrm{B}}] \quad , \tag{2}$$

A and B refer to different subpopulations, $\hat{a}_{i\mathrm{A}}$ and $\hat{b}_{i\mathrm{A}}$ are the discrimination and difficulty parameters for item i estimated for subpopulation A, and $\boldsymbol{\Gamma}_i^{-1}$ is the inverse of the asymptotic sampling variance-covariance matrix of $[\hat{b}_{i\mathrm{A}} - \hat{b}_{i\mathrm{B}}]$ and $[\hat{a}_{i\mathrm{A}} - \hat{a}_{i\mathrm{B}}]$. Lord (1980, p. 223) presented the formulas needed to evaluate Equation 1.

When person parameters are known, item parameters are estimated by the method of maximum likelihood, the null hypothesis of no bias is true, and the IRT model provides a reasonably good fit to the data, the distribution of Lord's bias statistic should be close to the chi-square distribution with two degrees of freedom. When person parameters are unknown and are estimated simultaneously with item parameters, the distributions of the item pa-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 2, June 1987, pp. 161–173

0146-6216/87/020161-13$1.90

rameter estimates and the chi-square statistic are unknown. The item parameter estimates are not necessarily consistent, in that the bias in the estimates has not been proven to decrease to 0 as the number of observations increases. Furthermore, the estimates are not necessarily asymptotically efficient, and they may not be normally distributed with the variance-covariance matrix given by Lord (1980). This problem of maximum likelihood estimation in situations where the number of parameters increases with sample size was first studied by Neyman and Scott (1948) and was discussed in the context of IRT by Hambleton and Swaminathan (1985, pp. 127–129).

Wingersky and Lord (1984) computed standard error estimates of item parameters when both item and person parameters are unknown. They explicitly treated item and person parameters as estimated simultaneously. Their expressions for standard errors should be more accurate than those obtained from the Lord (1980) formulas, in which θ estimates are treated as known person parameters. However, Wingersky and Lord's (1984) variance-covariance matrix may be inaccurate for the same reasons that the variance-covariance matrix presented by Lord may be inaccurate. In particular, the number of examinees may be viewed as increasing without limit, but the number of items is fixed and quite small. Consequently, consistent estimation of the person parameters is impossible.

One alternative that may circumvent the problems with simultaneous estimation is marginal maximum likelihood estimation (Bock & Lieberman, 1970), which removes the person parameters from the likelihood function. Drasgow (in preparation) found marginal estimation of item parameters to be far superior to simultaneous (item and person parameter) estimation for short tests. Nonetheless, simultaneous estimation continues to be widely used, particularly for tests with a moderate to large number of items.

In summary, a user of Lord's chi-square test may encounter difficulties in practice because the item parameters, when they are estimated simultaneously with person parameters, may not have the usual properties of maximum likelihood estimates. Consequently, Lord's item bias statistic may not follow the chi-square distribution well enough to allow valid tests of item bias.

Previous Research on Lord's Chi-Square Statistic

Lord (1977) presented an empirical check on the chi-square test of item bias, using data from the 85-item Scholastic Aptitude Test, Verbal section (SAT-V). He divided a mixed sample (Black and White) of 4,500 examinees at random into two groups. The null hypothesis of no difference in a and b parameters across the two random samples was tested. The number of significant chi-squares at each of 10 significance levels of equal width (.10) was reported, as well as the number significant at $\alpha = .05$. Given that the null hypothesis is true, four items significantly biased by chance at $\alpha = .05$ would be expected; three items were found to be significantly biased. This is a limited check on the chi-square test of significance, however, because parameter estimates from only two samples were obtained, so that each item was tested only once for equivalence across the samples. Lord suggested that repeated comparisons of item parameters across independent equivalent samples would better indicate the stability of the chi-square test.

Shepard, Camilli, and Williams (1984) also reported an empirical check on Lord's chi-square test. With a sample of 3,000 high school seniors, Shepard et al. checked for item bias on a 27-item mathematics achievement test and on a 32-item vocabulary test. Rather than split the total sample of Black and White examinees into two random samples as Lord (1977) did, they first tested for bias across two randomly split subsamples of Whites. They then tested for biased items across two randomly split subsamples of Blacks. For the vocabulary test, when the two White samples were compared, about 15% of the items were biased at a nominal alpha level of .05. The comparisons across the two Black samples were not reported. For the mathematics test the percentages of significant chi-squares for both the White and Black samples were close to the nominal alpha. The authors used the rejection rates observed in the within-population

comparisons to establish critical values for cross-population comparisons.

As in Lord's (1977) check on the chi-square, the rejection rates of Shepard et al. (1984) were based on only one chi-square test per item; several independent tests across equivalent samples are needed to obtain stable rejection rates. Furthermore, estimation is poorer for the extreme parameters; thus the number of significant chi-squares when the null hypothesis is true may vary as a function of the item parameters. Numerous independent tests for each of the items would allow assessment of the relation of item parameters and number of significant chi-squares when the null hypothesis is true.

If the distribution of the parameter estimates is close to the limiting distribution with the variance-covariance matrix given by Lord's (1980) formulas, then the chi-square test should indicate non-equivalent items only at chance levels. When standard error estimates are smaller than they should be in both groups compared, the null hypothesis of no difference in item parameters is more likely to be rejected. Depending on how much the standard errors are underestimated, the null hypothesis may be rejected more often than expected by chance. Knowledge of the base rate of rejection of the null hypothesis would be useful to researchers who use Lord's chi-square in selecting an appropriate alpha level for their chi-square tests.

The following simulation evaluated the chi-square test using 50 independent chi-square tests per item in each cell of the design. Significance rates obtained when item and person parameters were estimated simultaneously were compared to significance rates obtained when item parameters were estimated given the person parameters. In this way, the effects of simultaneous estimation of item and person parameters could be reliably determined. Significance rates for Lord's chi-square test are reported so that users of the test may set the critical values of their chi-square tests accordingly. In order to assess the accuracy of Lord's expression for asymptotic standard errors, the standard errors of parameter estimates across replications were computed and compared to standard errors estimated with Lord's expression. The asymptotic distributions of parameter estimates were examined by z tests and Kolmogorov-Smirnov tests. Finally, relations among the item parameters and the number of significant chi-squares were assessed with correlation and regression analyses.

Method

Design

Two sizes of examinee pools (250 and 1,000) were used, and responses were generated using both the two-parameter (2PL) and three-parameter (3PL) logistic models. The sample size of 1,000 was included and an item pool of 50 was used because Lord (1968) recommended a sample size of at least 1,000 and a test length of at least 50 items to obtain good estimates of the a parameters.

The item parameters used to generate the 3PL data were taken from Lord's (1968) analysis of the SAT-V. These values of the item parameters, shown in Table 2, are realistic because they were taken from a real test. For the 2PL model the cs were set equal to 0. The as and bs that were used for the 3PL model were also used for the 2PL model in order to allow straightforward comparisons across the simulated 3PL and 2PL tests. These item parameters for the 2PL test are less realistic than item parameters taken from a test modeled with 2PL item characteristic curves.

Procedure

The chi-square values were calculated for all 4 combinations of model and sample size in two ways. First, chi-squares were calculated as they would be in practice; item and person parameters were estimated simultaneously and chi-squares were calculated with $\hat{\theta}$s. Second, chi-squares were calculated with item parameters that were estimated given the true θs. A more detailed description of the procedures follows.

For each of the four combinations of logistic model (2PL and 3PL) and sample size (250 and 1,000), 100 independent samples of respondents and their responses to the 50 items were simulated. LOGIST (version 2b; Wood, Wingersky, & Lord, 1976) was used to estimate item and person parameters from

Table 1
Proportion of Significant Chi-squares for 3PL and 2PL Models Across All Samples and All Items

	N = 1000		N = 250	
α	θ Estimated	θ Known	θ Estimated	θ Known
3PL				
.0005	.0056	.0004	.0020	.0000
.001	.0096	.0004	.0032	.0000
.005	.0272	.0048	.0120	.0024
.01	.0400	.0084	.0216	.0044
.05	.1208	.0388	.0868	.0284
.10	.1976	.0812	.1556	.0652
2PL				
.0005	.0032	.0004	.0044	.0008
.001	.0064	.0004	.0056	.0012
.005	.0196	.0032	.0132	.0048
.01	.0360	.0076	.0284	.0084
.05	.1196	.0508	.0956	.0320
.10	.1972	.0980	.1680	.0692

each of the 100 samples, in each condition. Because c parameters are often indeterminate, Lord (1980, p. 217) recommended, for bias detection procedures, holding the $\hat{c}$s fixed at the values estimated from the two comparison samples combined. For the present study, if the proportion of correct responses to an item was less than .15, $\hat{c}$ was held fixed at .75 times the proportion correct. Otherwise, the $\hat{c}$s were held fixed at .15. Then, for each of the 50 independent pairs of samples, item and person parameter estimates were used to calculate a chi-square index of item bias for each of the 50 items. Then maximum likelihood estimates of item parameters were obtained by estimating the item parameters using the true θs. The standard errors and chi-squares were computed again, this time using the item parameters estimated given the true θs.

The proportions of significant chi-squares over all 50 items and 50 tests were computed at p-values of .0005, .001, .005, .01, .05, and .10, under all four combinations of sample size and logistic model. The proportions statistically significant at each nominal alpha level were compared across the two estimation procedures.

Standard errors of the item parameter estimates were calculated in two ways. First, standard errors were calculated by Lord's (1980) formulas, and will be referred to here as formula standard errors. In addition, standard deviations were estimated by the usual formula for the unbiased estimator using the 100 item parameter estimates, and are referred to here as empirical standard errors. The means of the formula standard errors were computed across samples for each item (for both the $\hat{a}$s and the $\hat{b}$s). The formula standard errors were compared to the empirical standard errors by dividing the means of the formula standard errors by the empirical standard errors, for each item. These ratios were then averaged across items.

Results

Chi-Square Tests

Table 1 shows proportions of significant chi-squares (calculated over all 50 items and all 50 comparisons) at alpha levels of .0005, .001, .005, .01, .05, and .10, for both the 3PL and 2PL models and sample sizes of 1,000 and 250. In the second

and fourth columns are proportions of significant chi-squares obtained when both item and person parameters were estimated simultaneously. In the third and fifth columns are the proportions significant for the true θs.

When both item and person parameters were estimated simultaneously, the proportions of significant chi-squares were greater than expected. For 3PL data when $N = 1{,}000$, observed proportions significant ranged from over 11 times ($\alpha = .0005$) to almost twice ($\alpha = .10$) the expected proportions significant. When $N = 250$, the observed proportions significant ranged from 4 times ($\alpha = .0005$) to over 1.5 times ($\alpha = .10$) the expected proportions significant. In this condition, 5% of the chi-squares were greater than 8.688, and 1% of the chi-squares were greater than 13.764.

Similar results were obtained when responses were generated by the 2PL model. When $N = 1{,}000$, the observed proportions of significant chi-squares ranged from over six times ($\alpha = .0005$) to almost twice ($\alpha = .10$) the expected proportions significant. When $N = 250$, the observed proportions ranged from almost 9 times ($\alpha = .0005$) to over 1.5 times ($\alpha = .10$) the expected proportions significant.

Comparison of the proportions significant under the two estimation procedures shows that simultaneous estimation of item and person parameters inflated the chi-square statistics. For both the 3PL and 2PL models and both sample sizes, the proportions of significant chi-squares when θs were known did not vary greatly from chance expectations.

Standard Errors

When person parameters were assumed unknown, the means of the formula standard errors were consistently smaller than the empirical standard errors, for both the $\hat{a}$s and the $\hat{b}$s. The means of the formula standard errors and the empirical standard errors for each item are shown in Table 2 for the 3PL, $N = 1{,}000$ condition. (The results for simultaneous estimation in the other three conditions were similar and are thus not presented.)

When the item parameters were estimated holding θs at their true values, the means of the formula standard errors were much closer to the empirical standard errors. These mean formula standard errors and empirical standard errors are shown in Table 3 for the 3PL, $N = 1{,}000$ condition. Again, results were similar for the other conditions. Notice that the empirical standard errors shown in Table 3 are smaller than those shown in Table 2. This indicates that the estimates obtained given the true θs were better than those obtained when θs were unknown.

To summarize the results in the four cells of the design, the means of the formula standard errors were compared to the empirical standard errors by dividing the means of the formula standard errors by the empirical standard errors, for each item. These ratios were then averaged across items, and are shown in Table 4, for the $\hat{a}$s and the $\hat{b}$s and for all four combinations of sample size ($N = 1{,}000$ and $N = 250$) and logistic model (3PL and 2PL). The first row shows the ratios under the condition of simultaneous estimation of item and person parameters; the second row shows the ratios obtained when the item parameters were estimated given the true θs.

When item parameters were estimated simultaneously with person parameters, the means of the formula standard errors of the $\hat{a}$s ranged from about 78% (2PL, $N = 1{,}000$) to about 93% (3PL, $N = 250$) of the average empirical standard errors. The means of the formula standard errors of the $\hat{b}$s ranged from about 77% (2PL, $N = 1{,}000$) to about 82% (3PL, $N = 1{,}000$ and 2PL, $N = 250$) of the size of the average empirical standard errors. The ratios were much closer to 1 when item parameters were estimated with known θs. When both item and person parameters were estimated simultaneously, the standard errors were sufficiently underestimated to change the distribution of the test statistic substantially and increase rejection rates.

Distributions of Item Parameter Estimates

To check for bias of item parameter estimates,

Table 2
Number of Significant Chi-squares at α =.05, Empirical SE's, and Mean Formula SE's for the 3PL Model and N=1000

Item	a	b	c	Number Sig.*	SE of â		SE of b̂	
					Empirical	Formula	Empirical	Formula
1	1.1	-.7	.20	13	.150	.101	.079	.064
2	.7	-.6	.20	5	.077	.065	.114	.094
3	.4	.1	.20	8	.157	.115	.057	.046
4	.9	.9	.16	9	.104	.093	.075	.075
5	1.2	.7	.12	9	.180	.136	.067	.052
6	1.6	1.1	.06	6	.226	.324	.064	.046
7	1.6	1.1	.06	5	.193	.329	.073	.046
8	1.6	-.1	.16	16	.281	.164	.052	.038
9	1.2	0.5	.20	6	.135	.097	.062	.055
10	2.0	1.6	.16	7	.324	.394	.101	.063
11	1.0	1.6	.13	2	.189	.183	.112	.100
12	1.5	1.7	.09	3	.323	.396	.085	.070
13	1.0	.7	.15	6	.112	.097	.084	.064
14	1.1	2.0	.06	6	.318	.313	.166	.116
15	1.1	2.4	.09	2	.462	.339	.318	.211
16	2.0	1.4	.11	2	.169	.394	.070	.052
17	1.7	1.3	.17	14	.366	.240	.078	.058
18	.5	-.6	.20	3	.065	.053	.150	.130
19	.9	1.6	.11	3	.196	.181	.119	.102
20	1.3	.4	.18	8	.145	.111	.061	.050
21	1.1	1.2	.05	6	.238	.232	.096	.062
22	1.2	1.1	.05	6	.279	.243	.077	.055
23	1.3	.2	.20	8	.155	.106	.059	.049
24	1.3	.2	.20	10	.174	.104	.057	.050
25	.5	-.8	.20	4	.060	.054	.160	.144
26	.7	.5	.20	1	.075	.065	.083	.081
27	.7	.5	.20	4	.086	.064	.074	.083
28	.4	-.4	.20	1	.049	.048	.142	.151
29	.4	-.4	.20	5	.051	.049	.163	.148
30	1.2	-.5	.20	12	.151	.107	.075	.055
31	.7	-1.0	.20	4	.085	.067	.135	.115
32	.7	-.2	.20	4	.075	.064	.084	.081
33	.7	-.2	.20	3	.073	.064	.075	.080
34	.5	.0	.20	5	.063	.053	.113	.106
35	.9	.5	.14	7	.110	.087	.070	.063
36	1.1	1.4	.04	2	.245	.284	.283	.065
37	1.2	-.6	.20	6	.156	.108	.072	.058
38	1.2	-.6	.20	8	.153	.109	.062	.057
39	.6	-.5	.20	2	.063	.058	.109	.104
40	1.6	.3	.18	10	.218	.140	.058	.043
41	1.1	.0	.20	12	.118	.088	.070	.056
42	1.5	2.0	.06	3	.375	.424	.122	.094
43	1.9	1.9	.11	0	.352	.478	.102	.082
44	.9	-.5	.20	4	.101	.078	.088	.072
45	.7	-.5	.20	6	.088	.065	.103	.091
46	1.4	1.6	.11	2	.326	.347	.089	.070
47	1.4	1.6	.11	5	.323	.338	.090	.069
48	1.0	1.7	.08	3	.303	.269	.114	.094
49	1.2	1.1	.15	3	.204	.145	.083	.065
50	1.2	1.1	.15	8	.192	.144	.085	.066

*α = .05

Table 3
Number of Significant Chi-squares at α = .05, Empirical SE's, and Mean Formula SE's for the 3PL Model and N=1000 (MLE's of item parameters given true thetas)

Item	a	b	Number Sig.*	SE of â		SE of b̂	
				Empirical	Formula	Empirical	Formula
1	1.1	-.7	5	.091	.090	.066	.066
2	.7	-.6	1	.062	.064	.098	.092
3	1.4	.1	4	.108	.107	.051	.048
4	.9	.9	2	.090	.094	.065	.076
5	1.2	.7	2	.133	.133	.053	.053
6	1.6	1.1	2	.219	.281	.052	.051
7	1.6	1.1	1	.214	.282	.052	.050
8	1.6	-.1	0	.125	.132	.039	.042
9	1.2	0.5	4	.110	.096	.052	.056
10	2.0	1.6	2	.342	.336	.075	.072
11	1.0	1.6	1	.154	.162	.102	.108
12	1.5	1.7	0	.208	.330	.078	.081
13	1.0	.7	3	.095	.098	.070	.074
14	1.1	2.0	1	.207	.266	.138	.129
15	1.1	2.4	0	.292	.290	.296	.241
16	2.0	1.4	0	.227	.358	.054	.058
17	1.7	1.3	3	.226	.207	.064	.063
18	.5	-.6	3	.056	.055	.133	.124
19	.9	1.6	1	.140	.166	.117	.111
20	1.3	.4	5	.121	.109	.050	.050
21	1.1	1.2	0	.156	.203	.076	.069
22	1.2	1.1	2	.197	.213	.069	.060
23	1.3	.2	4	.110	.102	.049	.050
24	1.3	.2	2	.114	.100	.048	.050
25	.5	-.8	3	.055	.055	.147	.138
26	.7	.5	2	.066	.066	.079	.081
27	.7	.5	0	.072	.065	.069	.082
28	.4	-.4	1	.044	.050	.128	.146
29	.4	-.4	2	.046	.050	.141	.142
30	1.2	-.5	3	.097	.095	.065	.058
31	.7	-1.0	2	.068	.066	.109	.111
32	.7	-.2	1	.062	.063	.077	.081
33	.7	-.2	2	.061	.064	.070	.079
34	.5	.0	2	.055	.054	.105	.104
35	.9	.5	7	.092	.088	.065	.063
36	1.1	1.4	0	.159	.239	.071	.074
37	1.2	-.6	1	.093	.095	.063	.061
38	1.2	-.6	0	.090	.094	.055	.061
39	.6	-.5	1	.054	.059	.097	.101
40	1.6	.3	2	.140	.129	.042	.044
41	1.1	.0	5	.095	.085	.059	.057
42	1.5	2.0	0	.294	.351	.108	.109
43	1.9	1.9	0	.333	.413	.091	.093
44	.9	-.5	3	.083	.075	.077	.072
45	.7	-.5	3	.070	.064	.090	.089
46	1.4	1.6	1	.216	.283	.080	.081
47	1.4	1.6	1	.221	.279	.078	.080
48	1.0	1.7	0	.225	.229	.105	.107
49	1.2	1.1	2	.149	.140	.070	.068
50	.12	1.1	5	.142	.139	.074	.068

*α = .05

Table 4
Mean Ratios of Mean Formula SE's to Empirical SE's

Model	θ	N=1000 a	N=1000 b	N=250 a	N=250 b
3PL	Estimated	0.912	0.817	0.932	0.782
3PL	Known	1.072	1.004	1.108	0.988
2PL	Estimated	0.775	0.768	0.845	0.817
2PL	Known	0.997	1.014	1.015	1.021

z tests were used to compare the means of the item parameter estimates to the true parameter values. Items 1, 8, 16, and 17 were chosen for the z tests because they had extreme parameters and/or high rejection rates. For each item, the standard error of the mean used in the z test was obtained from the empirical standard error by dividing by the square root of the number of replications. The resulting zs and their p values for the 3PL, N = 1,000 condition are shown in Table 5. The means of the $\hat{a}$s and the $\hat{b}$s were significantly different ($p < .05$) from the parameters for all four items, and thus it appears that the bias of the simultaneous maximum likelihood estimates is still substantial despite the large sample size.

In order to determine whether the item parameter estimates were normally distributed, the Kolmogorov-Smirnov (K-S) goodness-of-fit test was used to compare the distribution of the 3PL item parameter estimates to a normal distribution, for the N = 1,000 condition only. The distributions of the estimates were compared to the normal distribution using the mean estimates of the parameters and the empirical standard errors for the test distributions. When sample moments are used for the test distribution, the K-S test is conservative (Massey, 1951). The resulting K-S statistics (z_{KS}) and p values are shown in Table 5, for Items 1, 8, 16, and 17. Only one of the eight tests yielded a significant K-S statistic; the distribution of the $\hat{a}$s for Item 16 deviates far from the normal distribution ($z_{KS} = 4.36$, $p < .001$). The distributions of the $\hat{a}$s and $\hat{b}$s of Items 1, 8, and 17 and the $\hat{b}$s of Item 16 approximate the normal distribution.

Relation of Item Parameters to Chi-Square Tests

Estimation tends to be poorer for the extreme values of the item parameters. Hence the distribution of Lord's bias statistic may stray further from the chi-square when extreme item parameters are tested. The numbers of significant chi-squares ($\alpha = .05$) for each item were plotted, with the b

Table 5
Z-tests and Kolmogorov-Smirnov Goodness-of-Fit Tests
for 3PL Model, N=1000

Item	a	b	Number Sig.*	a z-test z	a z-test p	a K-S Test z_{KS}	a K-S Test p	b z-Test z	b z-Test p	b K-S Test z_{KS}	b K-S Test p
1	1.1	-.7	13	3.80	.00	1.02	.24	13.79	.00	.61	.84
8	1.6	-.1	16	10.29	.00	.88	.42	10.39	.00	.96	.32
16	2.0	1.4	2	24.79	.00	4.35	.00	5.57	.00	.95	.32
17	1.7	1.3	14	2.10	.04	.81	.43	4.10	.00	.53	.93

*α = .05

parameter values on the abscissa and the *a* parameter values on the ordinate. The plots are shown in Figure 1a for the 3PL, $N = 1{,}000$ condition, in Figure 1b for the 3PL, $N = 250$ condition, in Figure 2a for the 2PL, $N = 1{,}000$ condition, and in Figure 2b for the 2PL, $N = 250$ condition. Table 6 shows the correlations between the number of significant chi-squares and the item parameter values, for all four conditions, and the correlations between *a*, *b*, and $a \times b$. All of the correlations are significant ($p < .05$), except for the correlations between the numbers of significant chi-squares and the *a*s in the 3PL, $N = 250$ condition, and the *b*s in the 2PL, $N = 1{,}000$ condition. The *a* and *b* parameters were correlated ($r = .55$, $p < .001$).

Least-squares multiple regression was used to further explore the relation between the number of significant chi-squares and the item parameters. For

Figure 1
Number of Significant Chi-Squares, $p < .05$, for Each Item as a Function of *a* and *b*, When Item and Person Parameters Were Estimated Simultaneously: 3PL Model, $N = 1{,}000$ and $N = 250$

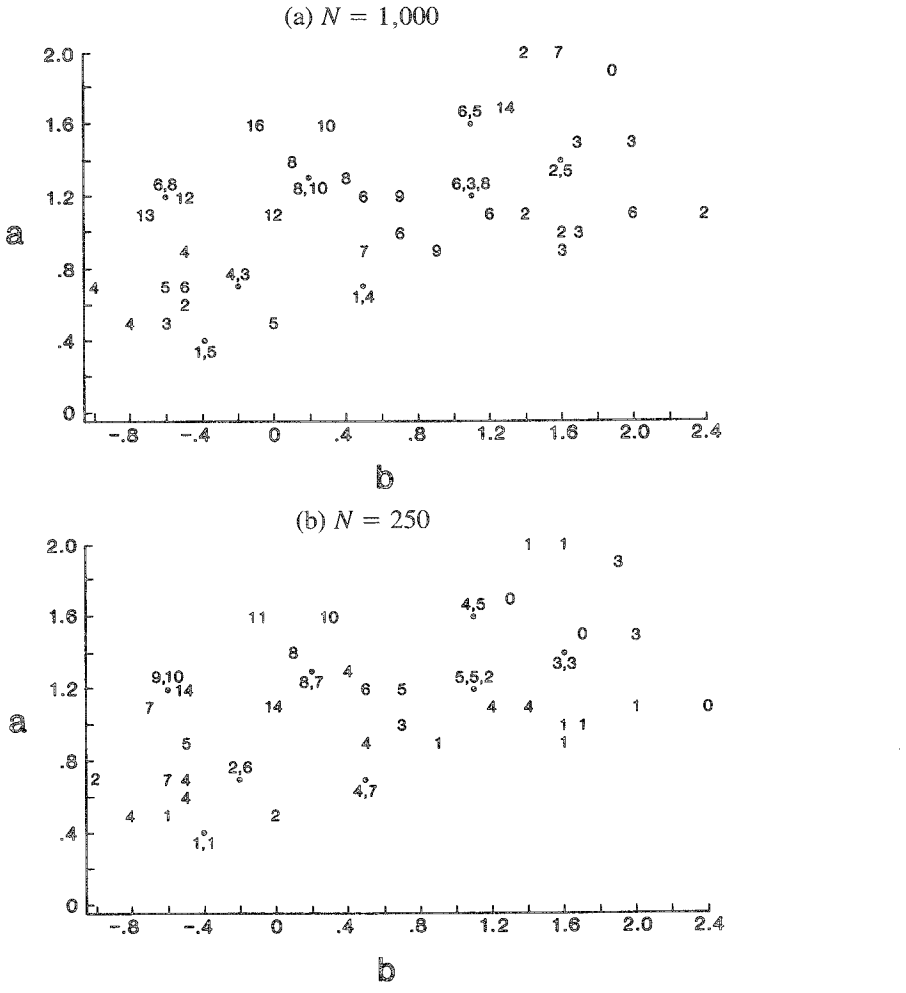


Table 6
Correlations Between a, b, and a x b, and Between a, b, a x b and the Number of Significant Chi-squares

			Number of Significant χ^2's			
			3PL		2PL	
	a	b	N=1000	N=250	N=1000	N=250
a	1.00	----	.28*	.12	.54***	.26*
b	.55***	1.00	-.27*	-.49***	.16	-.34**
a×b	.66***	.95***	-.30*	-.52***	.25*	-.36**

*p < .05, **p < .01, ***p < .001

each of the above four conditions, the number of significant chi-squares was regressed on *a*, *b*, and $a \times b$. The results are shown in Table 7. The multiple correlation coefficients were significant for all four conditions, ranging from .57 to .82. The largest regression coefficients for all four conditions were for the *a* parameters. Quadratic terms were also added to the regression equations, but they did not increase the *R*s significantly.

Discussion

Given the results presented above, researchers concerned with measurement equivalence may decide to (1) use the chi-square index and adjust the alpha level according to the results shown in Tables 1 through 3; (2) use the chi-square test without adjusting the alpha; or (3) use an alternative index. In deciding whether to use Lord's chi-square or an alternative item bias index, a researcher must weigh the advantages and disadvantages of using Lord's chi-square as a test of measurement equivalence, relative to alternative methods. The chi-square test is advantageous in that it represents a compelling conceptual definition of equivalence, and it is simple to calculate. Its main disadvantage is that the chi-square values are inflated due to underestimated standard errors, so that the probability of a Type I error is increased. Other indices of item equivalence and their advantages and disadvantages are discussed in Hulin, Drasgow, and Parsons (1983, pp. 152–183).

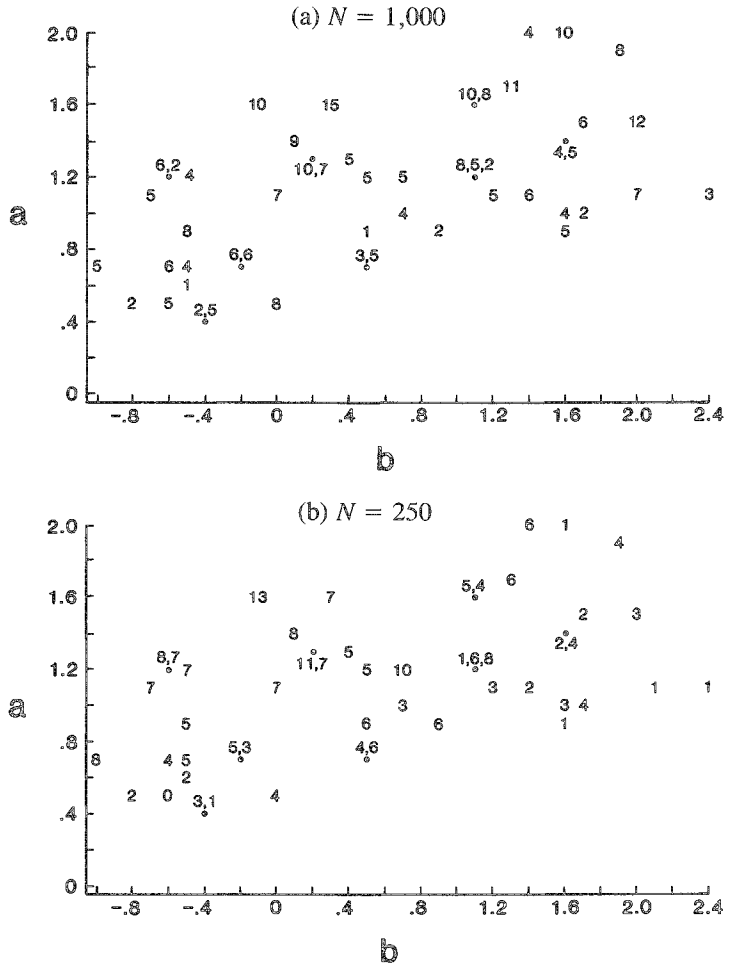
Researchers who intend to use the chi-square test may choose an alpha level based on the proportions significant found in this simulation. For example, a nominal alpha level of .001 yielded an actual rejection rate of slightly less than .01 for the 3PL model when θs were estimated and $N = 1{,}000$. An

Table 7
Multiple Correlations (R), Regression Coefficients (B), and F Statistics from the Regression of the Number of Significant Chi-squares (α = .05) on a, b, and a × b

Condition	R	B_a	B_b	B_{ab}	F
3PL					
N = 1000	.74	8.21	2.97	-4.90	18.23**
N = 250	.82	7.49	2.42	-4.95	30.82**
2PL					
N = 1000	.57	4.67	-.72	0.90	7.27*
N = 250	.77	6.53	1.97	-3.74	22.36**

*Significant at α = .01, **significant at α = .001.

Figure 2
Number of Significant Chi-Squares, $p < .05$, for Each Item as a Function of a and b, When Item and Person Parameters Were Estimated Simultaneously: 2PL Model, $N = 1{,}000$ and $N = 250$



actual rejection rate of this size is probably appropriate in light of the multiple comparisons usually made in an item bias study. Consequently, researchers with tests and sample sizes similar to those used in this simulation may want to choose a nominal alpha of .001 in order to achieve an actual rejection rate of approximately .01.

Whether the alpha level should be adjusted in a particular situation depends on the nature of the research. For example, when the test or scale has only a few items and test characteristic curves (TCCs) are found to be nearly equivalent, rejecting the null hypothesis when it is true would be particularly detrimental. In this case, a researcher would not wish to eliminate items on an already short test unless those items are truly biased. A more stringent nominal alpha level (based on the results shown in Tables 1 through 3) would reduce the number

of Type I errors to a more acceptable level, and would therefore reduce the chance of discarding truly unbiased items. On the other hand, when the test is composed of many items and the TCCs show substantial differences across groups, it may be more important to minimize Type II errors. In this case a nominal alpha level of .05 would help eliminate biased items, albeit at the cost of several Type I errors.

When the item parameters were estimated with the true θs, the proportions of significant chi-squares were close to the nominal alpha levels. These estimates seemed to have the properties usually associated with maximum likelihood estimates, and thus met the distributional assumptions required by the chi-square statistic. Also, the sampling variance-covariance matrix for the estimates was accurate enough so that the chi-square statistics were not inflated.

As shown by the plots in Figures 1 and 2 and by the regression analyses, the number of significant chi-squares is related to the item parameter values. The number of significant chi-squares should increase with parameter extremity because estimation tends to be poorer for the extreme values. This prediction holds for the *a* parameters, as evidenced by the moderate positive correlations between the *a* parameters and the numbers of significant chi-squares. Also, the regression coefficients for the *a* parameters are much higher than the coefficients for the *b* and $a \times b$ terms in the equations.

The results reported here may not be generalizable to conditions other than (1) normally distributed θs, (2) the particular set of SAT-V parameters, and (3) sample sizes of 250 and 1,000. The rejection rates obtained may not be applicable to conditions in practice that deviate greatly from the conditions established for this study. The effect of simultaneous estimation on the chi-square tests should decrease as the number of respondents decreases, because the formula standard errors increase.

Finally, marginal maximum likelihood estimates should perform better than the simultaneous estimates evaluated here. A computer simulation similar to the one presented here would determine whether the asymptotic properties of the parameter estimates are adequately approximated for sample sizes normally encountered. In addition, the formula standard errors need to be examined for their accuracy.

References

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

Drasgow, F. (in preparation). *An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model.* Champaign IL: University of Illinois, Department of Psychology.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications* (pp. 127–129). Boston: Kluwer-Nijhoff.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood IL: Dow Jones-Irwin.

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. Portinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam: Swets and Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (pp. 181–223). Hillsdale NJ: Erlbaum.

Massey, F. J., Jr. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association, 46,* 68–78.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16,* 1–32.

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

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST—A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum 76-6). Princeton NJ: Educational Testing Service.

Acknowledgments

The authors thank Charles L. Hulin for his comments on earlier versions of this manuscript, and David A. Harrison for his helpful suggestions and assistance with the statistical analyses.

Author's Address

Send requests for reprints or further information to Mary E. McLaughlin, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign IL 61820, U.S.A.