# An Investigation of Lord's Procedure for the Detection of Differential Item Functioning

Seock-Ho Kim and Allan S. Cohen, University of Wisconsin
Hae-Ok Kim, University of Nebraska

Type I error rates of Lord's $\chi^2$ test for differential item functioning were investigated using monte carlo simulations. Two- and three-parameter item response theory (IRT) models were used to generate 50-item tests for samples of 250 and 1,000 simulated examinees. Item parameters were estimated using two algorithms (marginal maximum likelihood estimation and marginal Bayesian estimation) for three IRT models (the three-parameter model, the three-parameter model with a fixed guessing parameter, and the two-parameter model). Proportions of significant $\chi^2$s at selected nominal $\alpha$ levels were compared to those from joint maximum likelihood estimation as reported by McLaughlin & Drasgow (1987). Type I error rates for the three-parameter model consistently exceeded theoretically expected values. Results for the three-parameter model with a fixed guessing parameter and for the two-parameter model were consistently lower than expected values at the $\alpha$ levels in this study. *Index terms: differential item functioning, item response theory, Lord's $\chi^2$.*

Lord (1977, 1980) proposed a $\chi^2$ statistic to test for differences in item parameters for the same item estimated in different groups under item response theory (IRT). When such differences are found, the items are said to be functioning differentially. More generally, an item functions differentially if the probability of a correct response is different for examinees at the same trait level who are from different groups (cf. Pine, 1977). Detection of such items is crucial because they threaten the validity of a test. The presence of differentially functioning items may also seriously interfere with efforts to equate different forms of a test.

The assumption of item parameter invariance under IRT (Baker, 1985) provides an ideal framework within which to examine differential item functioning (DIF). The basic building block of IRT is the item response function (IRF). The IRF describes the functional relationship between the probability of a correct response to an item and examinee trait level. The IRF is fully defined by its parameters. For example, for the three-parameter logistic model (3PLM), the item parameters include the item discrimination parameter ($a$), the item difficulty parameter ($b$), and the pseudo-guessing parameter ($c$); the person parameter is represented by trait level ($\theta$). The definition of DIF can be restated in terms of IRFs as follows:

> An item is said to be functioning differentially if IRFs obtained from different groups of examinees are different.

IRFs are identical if and only if the sets of item parameters calibrated in the different groups are equal.

Berk (1982), Holland & Wainer (1993), and Millsap & Everson (1993) reviewed classical test theory methods and IRT methods for detecting DIF. Using IRT, there are three primary approaches for DIF detection (e.g., Hambleton & Swaminathan, 1985; Ironson, 1983). The first approach compares item parameters estimated in two groups of examinees (Draba, 1977; Lord, 1977, 1980). The second approach measures the area between IRFs from two groups of examinees (Kim & Cohen, 1991; Linn, Levine, Hastings, & Wardrop, 1981; Raju, 1988, 1990; Rudner, 1977). The third approach assesses the fit of an item response model to the data (Linn & Harnisch, 1981; Wright, Mead, & Draba, 1976). The third approach also includes likelihood ratio tests to evaluate the significance of observed differences between item responses from two groups (Thissen,

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 18, No. 3, September 1994, pp. 217–228*

*0146-6216/94/030217-12$1.85*

1988, 1993). All three ap-
ınd to produce similar results,
ample sizes and longer tests—
ıe number of examinees is
:he number of items is greater
m, 1993; Kim & Cohen, in
& Slinde, 1993).
ısed the first approach to ex-
:scribed by Lord (1980). Pre-
's $\chi^2$, however, has indicated
parameter estimation condi-
I error control can become
ghlin & Drasgow, 1987).
gow's results were based on
)od estimation (JMLE) of item
ıe present study investigated
ent item parameter estimation
I error rate for Lord's $\chi^2$.

### re for DIF Detection

ysis, there are two groups of
'oup and the reference group.
s to the group of particular
ty group), and the reference
)up used as a basis of com-
. 217) described the follow-
letection:
meters for the two groups
.zing on the *b* estimates.
`c at the values obtained in
ı and *b* separately for each
ize again on the *b* estimates.
)are the *a* and *b* parameters
sing the $\chi^2$ statistic given by

IRT require that item param-
l on the same metric before
ade. Lord's standardization
one means of placing item
the same metric. Recent evi-
ts that the test characteristic
g metrics (Stocking & Lord,
curate than other currently
thods (Baker & Al-Karni,
1992). In the present study,
ıcteristic curve method was
used to place the estimates from the focal group onto the metric of the reference group.

Note that the second step in Lord's procedure does not include a test of the equality of *c* parameters when the 3PLM is used. If the 3PLM fits the data, then using the 3PLM suggests that all three item parameters should be tested simultaneously. It may be that the *c* parameter was not included by Lord (1980, p. 217) because estimates of this parameter are difficult to obtain using JMLE (cf. Thissen & Wainer, 1982).

### Extensions of the Procedure

The presence of DIF items may seriously affect the accuracy of efforts to link metrics, thereby resulting in spurious identification of DIF. Lord (1980, p. 220) discussed a purification method intended to reduce the potential effects of DIF items on item parameter estimates and subsequent DIF detection:

1. Analyze the test as described above.
2. Remove all items that have significantly different IRFs using the $\chi^2$ statistic. The remaining items now are considered a unidimensional pool, even when the groups are combined.
3. Combine the groups and estimate θs for each individual. These θs will be comparable.
4. For each group separately, estimate the *a* and *b* parameters while holding the θs fixed for all individuals at the values obtained in Step 3. Do this for all items, including those previously removed.
5. Compare the estimated item parameters using the $\chi^2$ statistic.

One problem with the purification approach is that it requires reestimation of item and θ parameters. Candell & Drasgow (1988) reported the following alternative procedure, due to Segall (1983), which is somewhat easier to implement:

1. Estimate item parameters independently in each group.
2. Link metrics across groups.
3. Calculate DIF indexes and remove DIF items.
4. Relink group metrics using only non-DIF items.
5. Recalculate DIF indexes and remove DIF items.

Steps 4 and 5 are continued until either no DIF items are detected or until the same set of DIF items is identified on subsequent iterations. This iterative

approach to linking has been found to be more accurate than the non-iterative approach (Candell & Drasgow, 1988; Cohen & Kim, 1993).

Lord (1980) indicated that the $\chi^2$ statistic is (1) asymptotic, (2) based on the assumption that $\theta$ parameters are known, and (3) applicable only with maximum likelihood estimates. Note that Assumption 2 cannot be achieved in a practical testing situation. Moreover, in a comparison of examinees based on certain group memberships (e.g., ethnicity) the number of examinees from the focal group may not be enough to satisfy Assumption 1.

### Parameter Estimation in Lord's Procedure

The estimation of item parameters for use in this procedure initially was based on JMLE. Recent developments in marginal maximum likelihood estimation (MMLE) (Bock & Aitkin, 1981) and marginal Bayesian estimation (MBE) (Mislevy, 1986), however, appear to have overcome some of the estimation problems of JMLE (e.g., Mislevy & Stocking, 1989). Therefore, a major concern of the present study was the level of Type I error control maintained in the application of Lord's procedure under MMLE and MBE.

McLaughlin & Drasgow (1987) found that JMLE of item and $\theta$ parameters resulted in Type I error rates that sometimes were seriously inflated over the nominal $\alpha$ level. MMLE and MBE differ from JMLE because item parameters are not estimated simultaneously with $\theta$ parameters. Further, MBE, as implemented in the computer program BILOG (Mislevy & Bock, 1990), has options for use of priors on item parameters, thereby providing Bayesian modal estimates for item parameters. Bayesian estimation methods have been found to avoid the tendency of maximum likelihood estimates to drift beyond a reasonable range of values (Mislevy, 1986). It is expected that with large sample sizes and long tests all three parameter estimation procedures—JMLE, MMLE, and MBE—will produce similar item parameter estimates. When small samples or short tests are used (e.g., when the number of examinees is less than 500 or the number of items is less than 20), however, an algorithm such as MBE is recommended (Mislevy & Stocking, 1989). Under such conditions, estimation method may play an important role in DIF detection with Lord's $\chi^2$ statistic.

Results using MMLE and MBE for the two-parameter logistic model (2PLM) showed that both MMLE and MBE provided more accurate parameter estimates and less inflated Type I error rates for Lord's $\chi^2$ statistic (Cohen & Kim, 1993; Lim & Drasgow, 1990) than JMLE. Results are lacking, however, for the applicability of Lord's (1980) $\chi^2$ for the 3PLM in the context of either MMLE or MBE (Millsap & Everson, 1993). The present study addressed this concern, replicating in part the study by McLaughlin & Drasgow (1987) using MMLE and MBE.

## Method

### Data Generation

Two sample sizes were used—$N = 1{,}000$ and $N = 250$. 1,000 simulated examinees were selected to represent a sample size appropriate for accurate estimation of item parameters for the 3PLM; 250 simulated examinees were selected to represent a sample size that was below the usually required minimum size of approximately 800 examinees with 20 items for the 3PLM (cf. Lord, 1968; Mislevy & Stocking, 1989; Swaminathan & Gifford, 1983).

Both the 3PLM,

$$P_j(\theta) = c_j + (1 - c_j)\frac{1}{1+\exp\left[-1.7a_j(\theta - b_j)\right]}, \qquad (1)$$

where $j$ designates an item, and the 2PLM,

$$P_j(\theta) = 1 + \frac{1}{1+\exp\left[-1.7a_j(\theta - b_j)\right]}, \qquad (2)$$

were used to generate datasets. For each of the four combinations of $N$ (1,000 and 250) and IRT models (3PLM and 2PLM), 100 replications of a 50-item simulated test were generated using the computer program GENIRV (Baker, 1988a). Thus, 400 datasets were generated.

The $\theta$s were sampled from the standard normal distribution, $\theta_i \sim N(0,1)$, where $i$ designates an examinee. Item parameter values used to generate the 3PLM data were taken from McLaughlin & Drasgow (1987) in order to permit a comparison between the results from that study and those from the present

study. Table 1 shows the generating item parameters; these values were originally reported by Lord (1968). The same $a$ and $b$ parameters were used to generate the datasets for the 2PLM.

### Item Parameter Estimation

Each dataset was analyzed with two estimation algorithms—MMLE and MBE—using the computer program BILOG (Mislevy & Bock, 1990). When no priors are used in BILOG, the resulting algorithm is MMLE. The BILOG default priors were used for MBE. For the 3PLM and the 2PLM, the prior distribution for the log item discrimination (log $a_j$) was normal with mean 0 and standard deviation .5. For the 3PLM, the prior distribution for $c_j$ was beta with hyperparameters 5 and 17.

Datasets generated with the 3PLM also were analyzed by the 3PLM with a fixed $c_j$ parameter (3PLM-$c$). The $c$s were fixed at .15, the average of the $c$ parameters, by imposing a prior distribution. For the 3PLM-$c$, item parameters were estimated under two estimation conditions: with (i.e., MBE) and without (i.e., MMLE) the normal prior on the log $a_j$s. Altogether, 1,200 sets of parameter estimates were obtained (i.e., 100 replications × two sample sizes × three IRT models × two estimation methods).

### Metric Transformation

50 pairs of simulated groups were formed randomly from the 100 replications of each of the 12 conditions (two sample sizes × three IRT models × two estimation methods). In each pair, one dataset was identified as the reference group and the other as the focal group. Item parameter estimates were equated onto the same metric using the test characteristic curve method of Stocking & Lord (1983) as implemented in EQUATE 2.0 (Baker, 1993). The equating coefficients, $A$ and $B$, then were used to place the item parameter estimates and the estimated variance (Var) and covariance (Cov) matrices from the focal group onto the metric of the reference group. The following equations were used:

$$\hat{a}_{jF}^{*} = \hat{a}_{jF}/A\,, \tag{3}$$

$$\hat{b}_{jF}^{*} = A \times \hat{b}_{jF} + B\,, \tag{4}$$

$$\mathrm{Var}\left(\hat{a}_{jF}^{*}\right) = \mathrm{Var}\left(\hat{a}_{jF}\right)/A^{2}\,, \tag{5}$$

$$\mathrm{Var}\left(\hat{b}_{jF}^{*}\right) = A^{2} \times \mathrm{Var}\left(\hat{b}_{jF}\right), \tag{6}$$

$$\mathrm{Cov}\left(\hat{a}_{jF}^{*}, \hat{c}_{jF}^{*}\right) = \mathrm{Cov}\left(\hat{a}_{jF}, \hat{c}_{jF}\right)/A\,, \tag{7}$$

and

$$\mathrm{Cov}\left(\hat{b}_{jF}^{*}, \hat{c}_{jF}^{*}\right) = A \times \mathrm{Cov}\left(\hat{b}_{jF}, \hat{c}_{jF}\right), \tag{8}$$

where ^ denotes an estimate, * denotes the transformed value, and the subscript F indicates that the estimate is obtained from the focal group. Note that for the 3PLM $\hat{c}_{jF}^{*} = \hat{c}_{jF}$. Only Equations 3–6 were used for the 3PLM-$c$ and the 2PLM.

### Lord's $\chi^2$

For each item, Lord's $\chi^2$ statistic for the 3PLM was obtained by

$$\chi_j^2 = \mathbf{v}_j' \boldsymbol{\Sigma}_j^{-1} \mathbf{v}_j\,, \tag{9}$$

where

$$\begin{aligned}\mathbf{v}_j &= \mathbf{v}_{jR} - \mathbf{v}_{jF}^{*} = \left(\hat{a}_{jR}, \hat{b}_{jR}, \hat{c}_{jR}\right)' - \left(\hat{a}_{jF}^{*}, \hat{b}_{jF}^{*}, \hat{c}_{jF}^{*}\right)' \\ &= \left(\hat{a}_{jR} - \hat{a}_{jF}^{*}, \hat{b}_{jR} - \hat{b}_{jF}^{*}, \hat{c}_{jR} - \hat{c}_{jF}^{*}\right)'\end{aligned} \tag{10}$$

and

$$\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_{jR} + \boldsymbol{\Sigma}_{jF}^{*}\,, \tag{11}$$

where $\boldsymbol{\Sigma}_{jR}$ is the estimated variance/covariance matrix of $\mathbf{v}_{jR}$, and $\boldsymbol{\Sigma}_{jF}^{*}$ is the (transformed) estimated variance/covariance matrix of $\mathbf{v}_{jF}^{*}$, and R indicates the reference group. For the 3PLM, $\mathbf{v}_j$ was a 3 × 1 vector, $\boldsymbol{\Sigma}_j$ was a 3 × 3 matrix, and $\chi^2$ had three degrees of freedom. For the 3PLM-$c$ and 2PLM, $\mathbf{v}_j$ was a 2 × 1 vector, $\boldsymbol{\Sigma}_j$ was a 2 × 2 matrix that involved both $a_j$ and $b_j$ estimates, and the $\chi^2$ had two degrees of freedom.

For each of the 50 items, 50 independent $\chi^2$s were calculated. The proportion of significant $\chi^2$s was calculated at nominal $\alpha$ levels of .0005, .001, .005, .01, .05, and .1. The first three steps of the iterative linking procedure of Candell & Drasgow (1988), described above, were used here to obtain Type I error rates.

**Table 1**
Number of Significant $\chi^2$s at $\alpha = .05$ for the 3PLM, the 3PLM-*c*, and the 2PLM

| | | | | 3PLM | | | | 3PLM-*c* | | | | 2PLM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Parameter | | | $N = 1{,}000$ | | $N = 250$ | | $N = 1{,}000$ | | $N = 250$ | | $N = 1{,}000$ | | $N = 250$ | |
| Item | $a_j$ | $b_j$ | $c_j$ | MBE | MMLE | MBE | MMLE | MBE | MMLE | MBE | MMLE | MBE | MMLE | MBE | MMLE |
| 1 | 1.1 | −.7 | .20 | 18 | 33 | 6 | 30 | 1 | 2 | 1 | 2 | 3 | 3 | 2 | 4 |
| 2 | .7 | −.6 | .20 | 16 | 42 | 0 | 29 | 3 | 3 | 0 | 0 | 4 | 4 | 0 | 0 |
| 3 | 1.4 | .1 | .20 | 20 | 27 | 6 | 29 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 4 | .9 | .9 | .16 | 9 | 18 | 8 | 26 | 0 | 0 | 2 | 3 | 5 | 5 | 1 | 1 |
| 5 | 1.2 | .7 | .12 | 8 | 20 | 2 | 11 | 3 | 4 | 1 | 1 | 5 | 5 | 1 | 3 |
| 6 | 1.6 | 1.1 | .06 | 4 | 6 | 4 | 11 | 0 | 0 | 0 | 4 | 3 | 3 | 1 | 1 |
| 7 | 1.6 | 1.1 | .06 | 8 | 11 | 3 | 7 | 3 | 3 | 1 | 1 | 4 | 5 | 3 | 3 |
| 8 | 1.6 | −.1 | .16 | 16 | 21 | 5 | 23 | 0 | 2 | 2 | 2 | 3 | 4 | 2 | 2 |
| 9 | 1.2 | .5 | .20 | 16 | 25 | 3 | 20 | 2 | 2 | 2 | 2 | 1 | 1 | 3 | 3 |
| 10 | 2.0 | 1.6 | .16 | 4 | 4 | 1 | 4 | 2 | 3 | 1 | 2 | 1 | 1 | 0 | 0 |
| 11 | 1.0 | 1.6 | .13 | 11 | 13 | 13 | 11 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| 12 | 1.5 | 1.7 | .09 | 8 | 10 | 5 | 11 | 0 | 0 | 1 | 2 | 1 | 1 | 0 | 0 |
| 13 | 1.0 | .7 | .15 | 8 | 19 | 10 | 21 | 1 | 1 | 2 | 3 | 5 | 6 | 3 | 3 |
| 14 | 1.1 | 2.0 | .06 | 12 | 15 | 2 | 3 | 0 | 0 | 0 | 0 | 3 | 3 | 0 | 0 |
| 15 | 1.1 | 2.4 | .09 | 6 | 10 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 2.0 | 1.4 | .11 | 3 | 6 | 3 | 8 | 2 | 2 | 0 | 3 | 0 | 0 | 0 | 0 |
| 17 | 1.7 | 1.3 | .17 | 6 | 7 | 1 | 5 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| 18 | .5 | −.6 | .20 | 14 | 34 | 5 | 30 | 6 | 6 | 1 | 1 | 3 | 3 | 3 | 3 |
| 19 | .9 | 1.6 | .11 | 14 | 20 | 4 | 4 | 3 | 3 | 1 | 1 | 0 | 0 | 1 | 1 |
| 20 | 1.3 | .4 | .18 | 19 | 24 | 7 | 23 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 0 |
| 21 | 1.1 | 1.2 | .05 | 6 | 6 | 3 | 6 | 2 | 2 | 1 | 3 | 1 | 1 | 1 | 1 |
| 22 | 1.2 | 1.1 | .05 | 8 | 13 | 4 | 13 | 1 | 2 | 0 | 0 | 3 | 3 | 1 | 1 |
| 23 | 1.3 | .2 | .20 | 21 | 29 | 6 | 24 | 0 | 0 | 4 | 5 | 5 | 5 | 1 | 2 |
| 24 | 1.3 | .2 | .20 | 16 | 27 | 5 | 24 | 2 | 2 | 3 | 3 | 3 | 4 | 0 | 2 |
| 25 | .5 | −.8 | .20 | 15 | 34 | 5 | 28 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 3 |
| 26 | .7 | .5 | .20 | 14 | 30 | 10 | 35 | 0 | 0 | 2 | 1 | 5 | 5 | 5 | 5 |
| 27 | .7 | .5 | .20 | 23 | 33 | 12 | 27 | 1 | 1 | 2 | 3 | 0 | 0 | 1 | 1 |
| 28 | .4 | −.4 | .20 | 16 | 38 | 6 | 22 | 1 | 1 | 3 | 2 | 1 | 1 | 0 | 0 |
| 29 | .4 | −.4 | .20 | 16 | 35 | 5 | 23 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 30 | 1.2 | −.5 | .20 | 20 | 33 | 2 | 30 | 2 | 3 | 1 | 4 | 3 | 3 | 0 | 1 |
| 31 | .7 | −1.0 | .20 | 17 | 37 | 7 | 27 | 0 | 0 | 1 | 1 | 2 | 2 | 0 | 0 |
| 32 | .7 | −.2 | .20 | 26 | 39 | 6 | 28 | 3 | 2 | 1 | 0 | 3 | 3 | 2 | 2 |
| 33 | .7 | −.2 | .20 | 20 | 39 | 7 | 33 | 1 | 1 | 1 | 3 | 0 | 0 | 1 | 1 |
| 34 | .5 | 0.0 | .20 | 18 | 36 | 11 | 31 | 3 | 3 | 3 | 3 | 1 | 1 | 2 | 3 |
| 35 | .9 | .5 | .14 | 9 | 26 | 5 | 17 | 3 | 2 | 1 | 3 | 3 | 3 | 2 | 2 |
| 36 | 1.1 | 1.4 | .04 | 10 | 13 | 2 | 5 | 3 | 3 | 0 | 2 | 0 | 0 | 0 | 0 |
| 37 | 1.2 | −.6 | .20 | 16 | 33 | 5 | 25 | 2 | 2 | 0 | 1 | 1 | 1 | 1 | 1 |
| 38 | 1.2 | −.6 | .20 | 21 | 38 | 7 | 29 | 0 | 2 | 1 | 2 | 3 | 3 | 0 | 1 |
| 39 | .6 | −.5 | .20 | 15 | 40 | 9 | 29 | 4 | 4 | 2 | 1 | 2 | 2 | 1 | 1 |
| 40 | 1.6 | .3 | .18 | 18 | 25 | 9 | 20 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 |
| 41 | 1.1 | 0.0 | .20 | 18 | 26 | 8 | 30 | 3 | 3 | 0 | 1 | 1 | 1 | 0 | 0 |
| 42 | 1.5 | 2.0 | .06 | 5 | 7 | 2 | 4 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 43 | 1.9 | 1.9 | .11 | 4 | 4 | 1 | 3 | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 2 |
| 44 | .9 | −.5 | .20 | 12 | 29 | 1 | 33 | 3 | 3 | 0 | 0 | 2 | 2 | 5 | 4 |
| 45 | .7 | −.5 | .20 | 18 | 42 | 2 | 30 | 0 | 0 | 2 | 3 | 2 | 2 | 1 | 1 |
| 46 | 1.4 | 1.6 | .11 | 5 | 7 | 5 | 8 | 1 | 1 | 0 | 0 | 2 | 2 | 1 | 1 |
| 47 | 1.4 | 1.6 | .11 | 8 | 11 | 2 | 6 | 1 | 2 | 0 | 0 | 2 | 2 | 0 | 0 |
| 48 | 1.0 | 1.7 | .08 | 12 | 14 | 6 | 8 | 1 | 1 | 0 | 1 | 2 | 2 | 0 | 0 |
| 49 | 1.2 | 1.1 | .15 | 11 | 12 | 7 | 8 | 2 | 2 | 0 | 2 | 3 | 3 | 4 | 4 |
| 50 | 1.2 | 1.1 | .15 | 7 | 11 | 7 | 13 | 3 | 4 | 1 | 2 | 1 | 1 | 1 | 1 |
| Total | | | | 645 | 1,132 | 256 | 927 | 76 | 85 | 49 | 81 | 100 | 104 | 58 | 67 |

## Results

### Item Parameters and Lord's $\chi^2$

Table 1 shows the number of significant $\chi^2$s for each item. There were large numbers of significant $\chi^2$s for the 3PLM. Item 1, for example, with $a_j = 1.1$, $b_j = -.7$, and $c_j = .20$, yielded 18 significant $\chi^2$s for the large sample ($N = 1{,}000$) MBE condition. Based on the 50 independent $\chi^2$s calculated for each item (from the 50 pairs of groups formed randomly from the 100 replications), the expected number of significant $\chi^2$s for one item at $\alpha = .05$ was 2.5; for all 50 items, the expected number was 125. The number of significant $\chi^2$s observed for this condition across replications and items was 645 or 5.16 times more than would be expected by chance. This same pattern was present for the 3PLM for both estimation methods at both sample sizes. Note that MBE had a smaller total than MMLE.

Table 1 also contains the number of significant $\chi^2$s for the 3PLM-*c*. For Item 1 with $N = 1{,}000$, there was one significant $\chi^2$ in the MBE condition; there were two for the $N = 1{,}000$, MMLE condition; one for the $N = 250$, MBE condition; and two for the $N = 250$, MMLE condition. Each of these was smaller than the 2.5 expected by chance at the $\alpha = .05$ level. Likewise, the number of significant $\chi^2$s observed over all 50 items was 76 for the $N = 1{,}000$, MBE condition and was less than expected by chance at the $\alpha = .05$ level. All four sample size × estimation method conditions yielded smaller numbers of significant $\chi^2$s for the 3PLM-*c* than expected by chance (i.e., 125). Table 1 also contains data for the 2PLM at the $\alpha = .05$ level. Results for both MBE and MMLE at each sample size yielded smaller numbers of significant $\chi^2$s than expected.

### Type I Error Rates

Table 2 presents the proportion of significant $\chi^2$s for the 12 conditions. The proportion of significant $\chi^2$s for MMLE was equal to or higher than that for MBE for the three IRT models under all except two conditions ($N = 1{,}000$, 3PLM-*c*, $\alpha = .001$ and .01). These proportions were always higher for the 3PLM than for the 3PLM-*c* or 2PLM.

With respect to Type I errors, Table 2 also shows

**Table 2**
Proportion of Significant $\chi^2$s for the 3PLM, 3PLM-*c*, and 2PLM Across All Samples and All Items

| Model and α | $N = 1{,}000$ MBE | $N = 1{,}000$ MMLE | $N = 250$ MBE | $N = 250$ MMLE |
|---|---|---|---|---|
| 3PLM | | | | |
| .0005 | .0924 | .2868 | .0008 | .2314 |
| .001 | .1064 | .2992 | .0096 | .2444 |
| .005 | .1460 | .3468 | .0316 | .2824 |
| .01 | .1652 | .3660 | .0432 | .2984 |
| .05 | .2580 | .4528 | .1024 | .3708 |
| .1 | .3160 | .4948 | .1572 | .4220 |
| 3PLM-*c* | | | | |
| .0005 | .0004 | .0004 | 0.0000 | .0004 |
| .001 | .0008 | .0004 | 0.0000 | .0008 |
| .005 | .0020 | .0020 | .0016 | .0024 |
| .01 | .0044 | .0040 | .0036 | .0052 |
| .05 | .0304 | .0340 | .0196 | .0324 |
| .1 | .0692 | .0796 | .0492 | .0696 |
| 2PLM | | | | |
| .0005 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| .001 | .0004 | .0004 | 0.0000 | 0.0000 |
| .005 | .0028 | .0032 | .0020 | .0020 |
| .01 | .0084 | .0092 | .0036 | .0084 |
| .05 | .0400 | .0416 | .0232 | .0268 |
| .1 | .0820 | .0864 | .0512 | .0600 |

that error rates were seriously inflated for the 3PLM for all α levels and were deflated for the 3PLM-*c* and the 2PLM. Loss of Type I error control for the 3PLM also was higher for MMLE in the $N = 1{,}000$ than in the $N = 250$ condition. The best control of Type I errors was found under this condition for both the 3PLM-*c* and the 2PLM. Finally, loss of Type I error control for the 3PLM for MMLE was markedly higher than for MBE at all α levels in this study.

For the 3PLM-*c*, the proportions of significant $\chi^2$s were all smaller than the respective α values. Lord's $\chi^2$ for the 3PLM-*c* appears to be somewhat conservative for the combinations analyzed in this study. At the .01 α level, observed proportions of significant $\chi^2$s were slightly less than half of that expected. The larger sample size showed generally better congruence with the expected Type I error rates, as did MMLE.

Results for the 2PLM were similar to those for the 3PLM-*c*. For $N = 1{,}000$, the obtained proportions of significant $\chi^2$s were only slightly smaller than the expected rates for both MBE and MMLE. For $N = 250$, the obtained proportions of signifi-

cant $\chi^2$s at the .05 and .1 $\alpha$ levels considered in this study were approximately half of what would be expected.

### Relationships Among Item Parameters and Significant $\chi^2$s

Spearman rank-order correlations among the item parameters and the number of significant $\chi^2$s are given in Tables 3, 4, and 5. The correlations among the item parameters were the same for all three IRT models and, therefore, are reported only in Table 3 (correlations with the $c_j$ parameter apply only to the 3PLM).

*3PLM.* The $a$ parameters had a positive correlation ($\rho = .500$) with the generating $b$ parameters. The correlation between the $c$ parameters in the 3PLM and the $a$ parameters was negative ($\rho = -.474$). Similarly, the correlation between the $b$ parameters and the $c$ parameters in the 3PLM was negative but much higher ($\rho = -.847$).

For the 3PLM, the correlations between $a_j$ and the number of significant $\chi^2$s from each of the sample size and estimation method conditions were all negative and moderate, ranging from −.350 to −.721 (see Table 3). Correlations between $b_j$ and the number of significant $\chi^2$s were negative and somewhat higher, except for the small sample MBE condition ($\rho = -.284$, $p > .05$). Correlations between $c_j$ and the number of significant $\chi^2$s were moderate to high, ranging from .360 to .846. The products of item parameters, $a_j \times b_j$ [studied by McLaughlin & Drasgow (1987)] and $b_j \times c_j$, were negatively correlated with the number of significant $\chi^2$s; correlations were significant for all but one of the sample size × estimation method conditions ($N = 250$, MBE).

For $N = 1{,}000$, the number of significant $\chi^2$s for MBE and MMLE were highly positively correlated ($\rho = .837$). The correlation between the number of significant $\chi^2$s from MBE and MMLE for $N = 250$ was .461. Moderate to high correlations were observed among the number of significant $\chi^2$s for all four combinations of sample size and estimation method, indicating consistency in identification of the same items as DIF items.

*3PLM-c.* Results for the 3PLM-$c$ are presented in Table 4. Significant correlations between item parameters and the number of significant $\chi^2$s were found only in the $N = 250$ MBE condition. The correlation between the number of significant $\chi^2$s from MBE and MMLE for $N = 1{,}000$ was .897; for $N = 250$, it was .582. The last column of Table 4 contains $\rho$s between the results of this study and the results reported by McLaughlin & Drasgow (1987)

**Table 3**
Spearman's $\rho$s Among Item Parameters and the Number of Significant $\chi^2$s for the 3PLM Estimated by MBE and MMLE

| Parameter, $N$, and Estimation Method | Parameter | | | | | | $N = 1{,}000$ | | $N = 250$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $a_j$ | $b_j$ | $c_j$ | $a_j \times b_j$ | $a_j \times c_j$ | $b_j \times c_j$ | MBE | MMLE | MBE | MMLE |
| $a_j$ | — | .500** | −.474** | .575** | .495** | .487** | −.428** | −.721** | −.350* | −.537** |
| $b_j$ | | — | −.847** | .969** | −.229 | .922** | −.717** | −.870** | −.284 | −.857** |
| $c_j$ | | | — | −.827** | .434** | −.660** | .722** | .840** | .360* | .846** |
| $a_j \times b_j$ | | | | — | −.172 | .914** | −.775** | −.902** | −.318* | −.862** |
| $a_j \times c_j$ | | | | | — | .016 | .253 | .023 | .066 | .227 |
| $b_j \times c_j$ | | | | | | — | −.666** | −.819** | −.162 | −.761** |
| $N = 1{,}000$ | | | | | | | | | | |
| MBE | | | | | | | — | .837** | .464** | .730** |
| MMLE | | | | | | | | — | .362* | .831** |
| $N = 250$ | | | | | | | | | | |
| MBE | | | | | | | | | — | .461** |
| MMLE | | | | | | | | | | — |

*Significant at $\alpha = .05$.
**Significant at $\alpha = .01$.

**Table 4**
Spearman's ρs Among Item Parameters and the Number of Significant $\chi^2$s for the 3PLM-*c* Estimated by MBE and MMLE

| Parameter, *N*, and Estimation | *N* = 1,000 | | *N* = 250 | | *N* = 1,000 |
|---|---|---|---|---|---|
| Model | MBE | MMLE | MBE | MMLE | JMLE |
| $a_j$ | −.203 | −.056 | −.313* | .040 | −.270 |
| $b_j$ | −.192 | −.243 | −.399** | −.160 | .312* |
| $a_j \times b_j$ | −.157 | −.220 | −.391** | −.137 | −.267 |
| *N* = 1,000 | | | | | |
| MBE | — | .897** | .024 | −.018 | .015 |
| MMLE | | — | .043 | .002 | .188 |
| *N* = 250 | | | | | |
| MBE | | | — | .582** | .223 |
| MMLE | | | | — | .333* |
| *N* = 1,000 | | | | | |
| JMLE | | | | | — |

*Significant at α = .05.
**Significant at α = .01.

for JMLE with *N* = 1,000. Only one correlation was significant between the number of significant $\chi^2$s from JMLE and the *N* = 250 condition for MBE and MMLE (.333 for the MMLE); none were significant for the *N* = 1,000 condition for MBE or MMLE. This suggests that the results from JMLE may not be consistent with those from either MBE or MMLE.

*2PLM.* Correlations in Table 5 between item parameters and the number of significant $\chi^2$s for the 2PLM were not significant for either MBE or MMLE for *N* = 1,000. For *N* = 250, the correlation between the *a* parameters and the number of significant $\chi^2$s for MBE (ρ = −.332) was significant. The *b*s and $a_j \times b_j$ were both negatively correlated with the number of significant $\chi^2$s in the small sample MMLE condition. A similar result was reported for the *N* = 1,000, JMLE condition by McLaughlin & Drasgow (1987). The ρs among the number of significant $\chi^2$s from the four sample size × estimation conditions in the present study were all significant. This indicates that MMLE and MBE results were more similar within sample size than between sample sizes. Correlations between the number of significant $\chi^2$s reported by McLaughlin and Drasgow and the *N* = 250 results from this study were significant. The results indicate that the patterns of significant $\chi^2$s were somewhat different for JMLE.

**Table 5**
Spearman's ρ Among Item Parameters and the Number of Significant $\chi^2$s for the 2PLM Estimated by MBE and MMLE, and JMLE for *N* = 1,000

| Parameter, *N*, and Estimation | *N* = 1,000 | | *N* = 250 | | *N* = 1,000 |
|---|---|---|---|---|---|
| Model | MBE | MMLE | MBE | MMLE | JMLE |
| $a_j$ | .017 | .040 | −.332* | −.199 | −.124 |
| $b_j$ | −.212 | −.210 | −.255 | −.301* | −.421** |
| $a_j \times b_j$ | −.214 | −.211 | −.249 | −.296* | −.367* |
| *N* = 1,000 | | | | | |
| MBE | | .996** | .333* | .480** | .207 |
| MMLE | | — | .335* | .489** | .192 |
| *N* = 250 | | | | | |
| MBE | | | — | .891** | .283* |
| MMLE | | | | — | .336* |
| *N* = 1,000 | | | | | |
| JMLE | | | | | — |

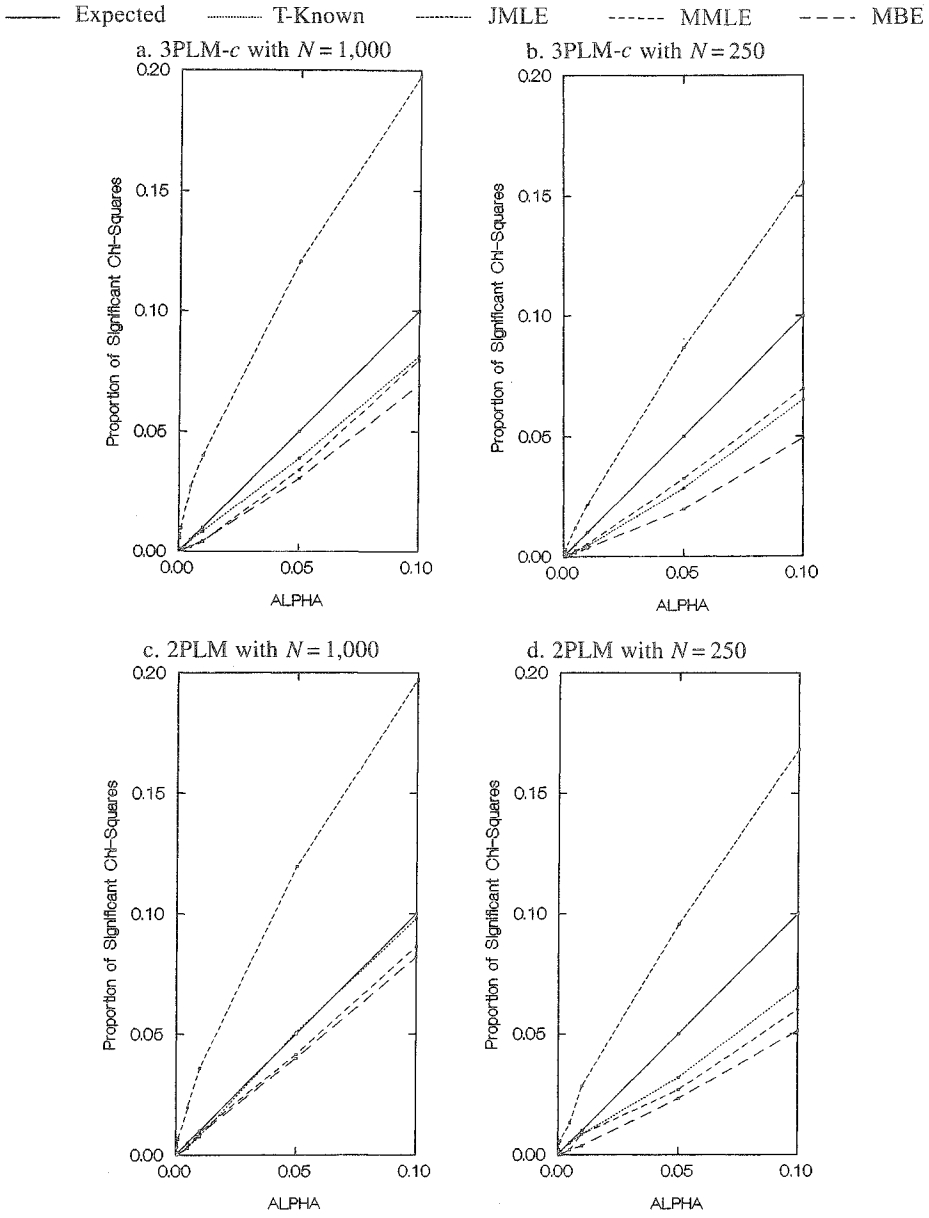
*Significant at α = .05.
**Significant at α = .01.

## Comparison of Type I Errors

Type I error rates for each estimation algorithm are plotted in Figures 1a–1d for the 3PLM-*c* and the 2PLM for both sample sizes. In addition, results for JMLE (i.e., θ unknown) and for θ level known (T-KNOWN) reported in Table 1 of McLaughlin & Drasgow (1987) are plotted. The solid line in each figure indicates the expected number of significant $\chi^2$s at each α level.

For the *N* = 1,000, 3PLM-*c* condition (Figure 1a), MBE, MMLE, and T-KNOWN yielded similar patterns in the proportions of significant $\chi^2$s. The plots of each of these proportions all lie below the expected value line, indicating that these estimation methods yielded somewhat deflated Type I error rates. Among the three estimation methods, MBE consistently yielded the lowest proportions of significant $\chi^2$s whereas T-KNOWN produced Type I error rates closest to the theoretically expected values. The JMLE results from McLaughlin & Drasgow (1987) yielded approximately twice the Type I error rates than would be expected at the α levels considered. Results for the 3PLM-*c* with *N* = 250 (Figure 1b) were similar to those for 3PLM-*c* with *N* = 1,000. MMLE yielded slightly better Type I error rates than

**Figure 1**
Proportion of Significant $\chi^2$s




T-KNOWN. MBE consistently yielded the lowest Type I error rates for all α levels and JMLE had slightly higher Type I error rates than expected by chance.

Figure 1c shows the results for the 2PLM, $N$ = 1,000 condition. Both MBE and MMLE yielded similar patterns of Type I error rates, slightly underestimating the expected levels. Type I error rates for T-KNOWN, however, were very close to rates at the nominal α levels. JMLE yielded Type I error rates that, in most instances, exceeded the expected lev-

els by a factor of two or more. Results for the 2PLM, $N = 250$ condition (Figure 1d) were similar to those for the 3PLM-$c$ condition (Figure 1b). MBE and MMLE consistently yielded Type I error rates that were approximately half that expected at all $\alpha$ levels. Results for T-KNOWN appeared to be slightly better than for MBE because they were closer to the expected rates. JMLE yielded error rates that were well above expected levels.

## Discussion

The primary purpose of this study was to examine Type I error rates for Lord's $\chi^2$ statistic under MMLE and MBE. The MBE and MMLE procedures implemented here were only one variation of each procedure. In fact, depending on the settings used, BILOG can yield a wide variety of MBE and MMLE models. Results for the 3PLM yielded error rates that were above those expected for the $\alpha$ levels used in this study. In some cases, error rates exceeded expectations by a factor of four or more. This is a serious problem. The 3PLM IRF is a function of all three parameters in the model. Any attempt to detect DIF in an item, therefore, should include all three parameters. If the 3PLM fits the data, then there is no a priori reason to automatically exclude one of the parameters from the comparison. Clearly, Lord's $\chi^2$ did not provide useful Type I error control for the 3PLM at either of the two sample sizes used in this study.

Type I error rates for the 3PLM-$c$ were all smaller than expected levels and clearly below those for the 3PLM. Based on these results, using Lord's $\chi^2$ with the 3PLM-$c$ provided a better model than the 3PLM with respect to Type I error control. In this study, the $c$ parameters were set to .15 for the 3PLM-$c$ even though the 3PLM and not the 3PLM-$c$ was used to generate data. In a practical testing situation, the $c$ parameters will not all be equal to a single value. In such a situation, one approach would be to follow Lord's (1980) suggestion and calculate estimates of the $c$ parameters from the combined (reference and focal) data. Fixing $\hat{c}_j$ in this way, the $a$ and $b$ parameters then can be estimated separately for the reference and focal groups. Setting $\hat{c}_j$ equal in the reference and focal groups, of course, reduces the test for DIF from three to two parameters.

The proportion of significant $\chi^2$s for the 2PLM were all less than the expected rates at each $\alpha$ level for the four sample size × estimation method conditions. Both MBE and MMLE for $N = 1{,}000$ yielded the best results; that is, the Type I error rates were closest to the expected numbers at the respective $\alpha$ levels. For $N = 250$, MBE yielded consistently smaller Type I error rates.

For all combinations of sample size and IRT models, MBE, MMLE, and T-KNOWN generally yielded Type I error results that were lower than theoretical expectations. MBE consistently yielded the lowest Type I error results, and T-KNOWN produced Type I errors closest to the theoretically expected rates. JMLE, however, consistently produced Type I error rates that far exceeded theoretical expectations [the T-KNOWN and JMLE results were from McLaughlin & Drasgow (1987)]. MBE and MMLE results for the 2PLM, $N = 1{,}000$ condition yielded the best Type I error rates.

The inflated Type I error rates that were reported for the 3PLM and the deflated Type I error rates reported for the 3PLM-$c$ and 2PLM may be due to problems in obtaining the estimated variance/covariance matrix. One way to examine this issue would be to calculate empirical standard errors as McLaughlin & Drasgow (1987) did. These then could be compared to values obtained from theoretical equations for the estimated variances and covariances of item parameter estimates under JMLE (Lord, 1980) and under the assumption that $\theta$ parameters are unknown but jointly estimated with item parameters (Wingersky & Lord, 1984).

For MMLE and MBE, however, no such simple equations are available unless the assumption of independence of items is made (Bock, Mislevy, & Thissen, 1991). In the absence of such equations, it seems reasonable to suggest that the deflated Type I error rates obtained for the 3PLM-$c$ and 2PLM were due to overestimated standard errors. Based on Holland (1990), the poor performance of Lord's $\chi^2$ statistic for the 3PLM may be partially due to the fact that $c_j$ is empirically underidentified in most cases, which may cause inflated values for the elements of the inverse of the estimated variance/covariance matrices. The result is an inflated test

statistic. Thissen & Wainer (1982) and Baker (1988b) also have noted similar problems for small samples in estimating the variance/covariance matrices used in Lord's $\chi^2$, especially for the 3PLM.

When actual test data are used, linking metrics in the context of detection of potential DIF requires attention to issues that can be controlled in simulations. Some of these issues using real data are: (1) the question of fit of an IRT model to the test data needs to be addressed, (2) differences between the $\theta$ distributions of the reference and focal groups may be present, and (3) the presence of DIF items may require some method such as iterative linking of metrics to improve the linking transformation and subsequent DIF detection (Candell & Drasgow, 1988). Lord's $\chi^2$ is an asymptotic statistic and consequently requires large sample sizes. Results from this study support this: The larger sample size ($N$ = 1,000) yielded better Type I error rates for both the 3PLM-$c$ and the 2PLM. In addition, Lord's $\chi^2$ is applicable to only maximum likelihood estimates. This also is in agreement with results from the present study as Type I error rates for item parameter estimates using MMLE were closer to theoretical expectation. The differences, however, between MBE and MMLE were generally minor.

Lord's $\chi^2$ is flexible because it can be applied to situations in which multiple groups are tested for DIF simultaneously. It also can be applied easily to DIF detection with graded response and other IRT models. Type I error rates with these other models also should be investigated. Lord's $\chi^2$ method also should be compared with the likelihood ratio test (Thissen et al., 1988, 1993), because the likelihood ratio test does not rely directly on the estimation of variance/covariance matrices. Finally, the statistical power of Lord's $\chi^2$ statistic is also an important issue that was not addressed here. The issue has been studied for the 2PLM, however, by Cohen & Kim (1993) and Lim & Drasgow (1990).

## References

Baker, F. B. (1985). *The basics of item response theory.* Portsmouth NH: Heinemann.

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

Baker, F. B. (1988b). The item log-likelihood surface for two- and three-parameter item characteristic curve models. *Applied Psychological Measurement, 12,* 387–395.

Baker, F. B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17,* 20.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28,* 147–162.

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias.* Baltimore MD: Johns Hopkins University Press.

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

Bock, R. D., Mislevy, R. J., & Thissen, D. (1991). *Item response theory.* Unpublished manuscript.

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 $\chi^2$ 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, Education Statistics Laboratory.

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

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601.

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

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 155–174). Vancouver, Canada: Educational Research Institute of British Columbia.

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.

Kim, S.-H., & Cohen, A. S. (in press). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item func-

tioning. *Applied Measurement in Education.*

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.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109–118.

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. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

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. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57–75.

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 Measurement, 14,* 197–207.

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational and Psychological Measurement, 53,* 301–314.

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. (ERIC Document Reproduction Service No. ED 137 337)

Segall, D. O. (1983). *Test characteristic curves, item bias and transformation to a common metric in item response theory: A methodological artifact with serious consequences and a simple solution.* Unpublished manuscript, University of Illinois, Department of Psychology, Champaign-Urbana.

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

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 13–30). New York: Academic Press.

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.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397–412.

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.

Wright, B. D., Mead, R., & Draba, R. (1976). *Detecting and correcting test item bias with a logistic response model* (Research Memorandum No. 22). Chicago: The University of Chicago, Department of Education, Statistical Laboratory.

## Acknowledgments

*Portions of this paper were presented at the annual meeting of the National Council on Measurement in Education, New Orleans LA, April, 1994.*

## Author's Address

Send requests for reprints or further information to Seock-Ho Kim, Testing and Evaluation, 1025 West Johnson Street, Madison WI 53706, U.S.A. Internet: shkim@tne.edsci.wisc.edu.