DOCUMENT RESUME

ED 420 690 TM 028 365

AUTHOR	Cohen, Allan S.; Kim, Seock-Ho; Wollack, James A.
TITLE	A Comparison of Item Response Theory and Observed Score DIF Detection Measures for the Graded Response Model.
PUB DATE	1998-04-14
NOTE	40p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 12-16, 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Identification; *Item Bias; Item Response Theory; *Scores; Test Construction; *Test Items
IDENTIFIERS	*Graded Response Model; Item Bias Detection

ABSTRACT

This paper provides a review of procedures for detection of differential item functioning (DIF) for item response theory (IRT) and observed score methods for the graded response model. In addition, data from a test anxiety scale were analyzed to examine the congruence among these procedures. Data from Nasser, Takahashi, and Benson (1997) were reanalyzed for purposes of this study. The data were obtained from participants' responses to an Arabic Version of Sarason's (1984) Reactions to Test (RTT) scale. The sample consisted of 421 tenth graders from two Arab high schools in the central district of Israel. Results indicated stronger agreement within IRT methods and within observed score methods than between these two sets of DIF detection methods. A discussion is included focusing on reasons for these similarities and differences. Results of this study can provide useful information about the relationships to expect between various DIF detection methods. (Contains 10 tables and 49 references.) (Author/SLD)

A Comparison of Item Response Theory And Observed Score DIF Detection Measures For the Graded Response Model

Allan S. Cohen
University of Wisconsin–Madison
Seock-Ho Kim
The University of Georgia
James A. Wollack
University of Wisconsin–Madison

April 14, 1998
Running Head: COMPARISON OF DIF DETECTION

Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, California

A Comparison of Item Response Theory And Observed Score DIF Detection Measures For the Graded Response Model

Abstract

This paper provides a review of procedures for detection of differential item functioning (DIF) for item response theory (IRT) and observed score methods for the graded response model. In addition, data from a test anxiety scale were analyzed to examine the congruence among these procedures. Results indicated stronger agreement within IRT methods and within observed score methods than between these two sets of DIF detection methods. A discussion is included focusing on reasons for these similarities and differences.

Key words: area measures, chi-square test, differential item functioning, generalized Mantel-Haenszel test, graded response model, item response theory, likelihood ratio test, Mantel test, simultaneous item bias test.

Introduction

Graded response items are particularly useful for test items in which examinees answers are not simply scored as correct or incorrect. As on any test, however, items which function differently in different groups need to be detected and, if necessary, removed, because they present a threat to the validity of the test. Although a number of methods for detection of such items have been developed for graded response items, either based on item response theory (IRT) or based on observed scores, very little research has compared results from these two sets of methods.

One problem which faces developers of tests using polytomous response items is that the different DIF detection indices tend to identify different sets of items on the test as functioning differentially (e.g., Ankenmann, Witt, & Dunbar, 1996; Chang, Mazzeo, & Roussos, 1996; Kim, Cohen, & Baker, 1996; Welch & Hoover, 1993; Zwick, Donoghue, & Grima, 1993). Given such a scenario, it can be difficult for one to determine which, if any, DIF indices should be used. In this paper, we provide a review of IRT and observed score methods for detecting DIF in graded response items with an eye toward examining what is measured by each index. We then provide a comparison of the procedures reviewed using a set of graded response test data.

The Graded Response Model and DIF

In the context of dichotomous IRT models, an item is said to be functioning differentially, when the probability of a correct response to the item is different for examinees at the same ability level but from different groups (Pine, 1977). The presence of such items on a test indicates that examinees at the same underlying θ may exhibit systematically different patterns of item responses. In this section, we describe the graded response model under IRT (Samejima, 1969, 1972) and methods for detecting DIF items in that model.

The item response function (IRF) is the basic building block of IRT. For a dichotomously scored item, the IRF is usually taken to refer to that function which characterizes the relationship between the probability of a correct response to an item and examinee trait level θ. There are, however, two IRFs for a dichotomous item, one for the correct response and one for the incorrect response.

The Graded Response Model

Samejima (1969, 1972) proposed a graded response model under IRT in which the category response function, $P_{jk}(\theta)$, describes the probability of response k to item j as a function of θ. For an item with K_j categories, $P_{jk}(\theta)$ is defined as

$$P_{jk}(\theta) = \begin{cases} 1 - P^*_{j1}(\theta) & \text{when } k = 1 \\ P^*_{j(k-1)}(\theta) - P^*_{jk}(\theta) & \text{when } k = 2, \ldots, (K_j - 1) \\ P^*_{j(K_j-1)}(\theta) & \text{when } k = K_j, \end{cases} \tag{1}$$

where $k = 1, \ldots, (K_j - 1)$. In Equation 1, $P^*_{jk}(\theta)$ is the cumulative category response function given by

$$P^*_{jk}(\theta) = \{1 + \exp[-\alpha_j(\theta - \beta_{jk})]\}^{-1}, \tag{2}$$

where α_j is the discrimination parameter for item j, β_{jk} is the location parameter of response category k for item j, and θ is the trait level parameter. The logistic model in Equation 2 is a homogeneous case of the general graded response model (Samejima, 1972, 1997). With $P^*_{j0}(\theta) = 1$ and $P^*_{jK_j}(\theta) = 0$, the category response function can be succinctly written as

$$P_{jk}(\theta) = P_{j(k-1)}(\theta) - P^*_{jk}(\theta). \tag{3}$$

Item True Score Functions. For a polytomously scored item such as the graded response item (Samejima, 1969, 1972), the item true score function describes the relationship between the expected value of the item score and examinee trait level.

Baker (1992) defined the true score function for the graded response model as

$$T(\theta) = \sum_{j=1}^{J} \sum_{k=1}^{K_j} u_{jk} P_{jk}(\theta), \tag{4}$$

where J is the number of items in the test and u_{jk} is the weight for response category k of item j. Weights are typically, but not necessarily, taken to be the same as the category values. For example, the weight for category 1 would be 1, and for category 4 it would be 4.

The item true score function for a single item j can be defined as

$$T_j(\theta) = \sum_{k=1}^{K_j} u_{jk} P_{jk}(\theta). \tag{5}$$

Definition of DIF. In the typical DIF study, there are two groups of examinees, the reference group and the focal group. For a dichotomous item under IRT, the IRF is the item true score function. For both dichotomous and graded response items, an item is considered to be functioning differentially, when the item true score functions in the reference and focal groups are not equal (Cohen, Kim, & Baker, 1993). That is, a DIF item is identified, when $T_{jR}(\theta) \neq T_{jF}(\theta)$. Further, the item true score functions from the reference and focal groups are identical if and only if the cumulative category response functions for the reference and focal groups are equal or the sets of item parameters from the reference and focal groups are equal. These two conditions are essentially equivalent.

Detection of DIF. The equality of sets of item parameters for graded response items can be tested using several different approaches. One approach, the chi-square test, is to compare item parameters estimated from the two groups (e.g., Cohen et al., 1993; Millsap & Everson, 1993). A second approach is to obtain and test area measures or distances between item true score functions (e.g., Cohen et al., 1993; Flowers, Oshima, & Raju, 1995; Raju, van der Linden, & Fleer, 1995). A third approach, the likelihood ratio (LR) test (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993; Wainer, Sireci, & Thissen, 1991), uses a likelihood ratio (Neyman & Pearson, 1928) to compare likelihood functions estimated from different groups in order to evaluate differences between item responses from the two groups. Thissen, Steinberg, and Wainer (1988) noted that the third approach is preferable for theoretical reasons because the first and second approaches may require estimates of variances and covariances of the item parameters. At the present time, computational difficulties impede obtaining accurate estimates of these variances and covariances.

Ankenmann et al. (1996) compared power and Type I error rates of the LR test and the Mantel (1963) test for DIF detection under the graded response model (Samejima, 1969, 1972). Ankenmann et al. (1996) used combined dichotomous and graded response item data and obtained the power and Type I error rates for a single studied graded response item in each data set under different sample sizes and ability conditions. The LR test was found to yield better power and control of Type I error than the Mantel procedure (Ankenmann et al., 1996). Kim and Cohen (in press) reported Type I error rates of the LR test for DIF detection for a graded response model with five ordered categories. Data were generated for a 30-item test for six combinations of sample sizes by underlying ability conditions. Type I error rates of the LR test were found to be within theoretically expected values at each of the nominal alpha levels considered. Analysis of Type I error rates for the chi-square test and the area measures described by Cohen et al. (1993), however, indicated mixed results

(Kim et al., 1996). Type I error control was conservative for the chi-square and the signed area measure but poor for the unsigned area measure. The LR test of DIF under the graded response model seems promising, but it can be computationally quite intensive (Thissen et al., 1993).

IRT Methods for DIF Detection

The Chi-Square Test. A χ^2 originally described by Lord (1980) for dichotomous IRT models can also be used to test the hypothesis that the parameters estimated for a graded response item are the same between reference and focal groups (Cohen et al. 1993). The χ^2 statistic for the graded response model item with K_j categories is computed as

$$\chi_j^2 = \hat{\xi}_j' \hat{\Sigma}_j^{-1} \hat{\xi}_j, \tag{6}$$

where $\hat{\xi}$ is the vector of difference between parameter estimates (i.e., $\hat{\xi}_j = \hat{\xi}_{jF} - \hat{\xi}_{jR}$) and $\hat{\Sigma}_j^{-1}$ is the inverse of the variance- covariance matrix, $\hat{\xi}_j$ (i.e., $\hat{\Sigma}_j = \hat{\Sigma}_{jR} + \hat{\Sigma}_{jF}$).

The vector of item parameter estimates for the reference group can be written as

$$\hat{\xi}_{jR} = \left[a_{jR}, b_{j1R}, \ldots, b_{j(K_{j-1})R}\right]' \tag{7}$$

and the variance-covariance matrix can be written as

$$\hat{\Sigma}_{jR} = \begin{bmatrix} \mathrm{Var}(a_{jR}) & \mathrm{Cov}(a_{jR}, b_{j1R}) & \cdots & \mathrm{Cov}(a_{jR}, b_{j(K_j-1)R}) \\ & \mathrm{Var}(b_{j1R}) & \cdots & \mathrm{Cov}(b_{j1R}, b_{j(K_j-1)R}) \\ & & \ddots & \vdots \\ & & & \mathrm{Var}(b_{j(K_j-1)R}) \end{bmatrix}. \tag{8}$$

The vector of item parameter estimates and the estimated variance-covariance matrix for the focal group can be defined similarly. There are K_j degrees of freedom for this extension of Lord's χ^2 for a graded response model with K_j categories.

The Signed Area. Raju (1988, 1990) developed a test of the signed area between item response functions for dichotomous models. An extension of this test for graded response items (Cohen et al., 1993) is given below. Let

$$\hat{T}_{jR}(\theta) = \sum_{k=1}^{K_j} u_{jk} \hat{P}_{jkR}(\theta) \tag{9}$$

be the estimate of the item true score function for item j in the reference group and let

$$\hat{T}_{jF}(\theta) = \sum_{k=1}^{K_j} u_{jk} \hat{P}_{jkF}(\theta) \tag{10}$$

be the estimate of the item true score function in the reference group, where $\hat{P}_{jkR}(\theta)$ and $\hat{P}_{jkF}(\theta)$ are the estimates of the cumulative category response functions for the reference and focal groups, respectively.

According to Cohen et al. (1993), the signed area S_j between the two item true score functions is obtained as

$$S_j = \int_{-\infty}^{\infty} \left[\hat{T}_{jR}(\theta) - \hat{T}_{jF}(\theta)\right] d\theta = \sum_{k=1}^{K_j-1} \left[u_{j(k+1)} - u_{jk}\right] (b_{jkF} - b_{jkR}), \tag{11}$$

where b_{jkR} and b_{jkF} are the estimates of β_{jkR} and β_{jkF}, respectively. The estimated variance of S_j is defined as

$$\begin{aligned} \mathrm{Var}(S_j) \quad = \quad & \sum_{k=1}^{K_j-1} \left[u_{j(k+1)} - u_{jk}\right]^2 \mathrm{Var}(b_{jkF}) + \\ & \sum_{k=1}^{K_j-1} \sum_{l=1}^{K_j-1} \left[u_{j(k+1)} - u_{jk}\right] \left[u_{j(l+1)} - u_{jl}\right] \mathrm{Cov}(b_{jkF}, b_{jlF}) + \\ & \sum_{k=1}^{K_j-1} \left[u_{j(k+1)} - u_{jk}\right]^2 \mathrm{Var}(b_{jkR}) + \\ & \sum_{k=1}^{K_j-1} \sum_{l=1}^{K_j-1} \left[u_{j(k+1)} - u_{jk}\right] \left[u_{j(l+1)} - u_{jl}\right] \mathrm{Cov}(b_{jkR}, b_{jlR}), \end{aligned} \tag{12}$$

where $k \neq l$.

The test statistics $Z(S_j)$ can be written as

$$Z(S_j) = \frac{S_j}{\sqrt{\mathrm{Var}(S_j)}} \tag{13}$$

and is based on the assumption that the observed signed areas S_j are normally distributed with mean 0 and variance given in Equation 12.

The Unsigned Area. Raju (1988, 1990) also developed an unsigned area test for the difference between item response functions for dichotomous items. Cohen et al. (1993) showed that the unsigned area, U_j, between the two item true score functions is obtained as

$$U_j = \int_{-\infty}^{\infty} \left|\hat{T}_{jR}(\theta) - \hat{T}_{jF}(\theta)\right| d\theta. \tag{14}$$

Expressing U_j in terms of cumulative category response functions gives

$$U_j = \int_{-\infty}^{\infty} \left| \sum_{k=1}^{K_j-1} \left[u_{j(k+1)} - u_{jk}\right] \left[\hat{P}^*_{jkR}(\theta) - \hat{P}^*_{jkR}(\theta)\right] \right| d\theta. \tag{15}$$

If either $\hat{T}_{jR}(\theta) \geq \hat{T}_{jF}(\theta)$ or $\hat{T}_{jR}(\theta) \leq \hat{T}_{jF}(\theta)$ for all θ, then

$$U_j = |S_j| . \tag{16}$$

Assuming that the S_j are normally distributed with mean 0 and variance as given in Equation 12, the expected value of U_j is

$$E(U_j) = \sqrt{\frac{2}{\pi}\mathrm{Var}(S_j)} \tag{17}$$

and the variance of U_j is

$$\mathrm{Var}(U_j) = \mathrm{Var}(S_j)\left(1 - \frac{2}{\pi}\right) \tag{18}$$

(Hogg & Craig, 1978). It should be noted that the assumption of normality for U_j may not be justified (Raju, 1990). Equation 13 also provides a test of the null hypothesis of no DIF only if either $\hat{T}_{jR}(\theta) \geq \hat{T}_{jF}(\theta)$ or $\hat{T}_{jR}(\theta) \leq \hat{T}_{jF}(\theta)$ for all θ. If either condition $\hat{T}_{jR}(\theta) \geq \hat{T}_{jF}(\theta)$ or $\hat{T}_{jR}(\theta) \leq \hat{T}_{jF}(\theta)$ for all θ does not hold, U_j may not have a closed form. In such a case, no statistical test is yet available for the null hypothesis. Even so, it still may be of interest to examine the size of U_j and to test its significance with the variance given in Equation 18.

The following approximation may be used for U_j: Select two points θ_L and θ_U such that $\theta_L \leq \theta_U$ and divide the range into N intervals. The area U_j then is estimated using the trapezoidal approximation of the bounded unsigned area (Burden & Faires, 1985) as

$$\begin{aligned} U_j \;=\; & \sum_{i=1}^{N} \left| \left[u_{j(k+1)} - u_{jk} \right] \left[\hat{P}^{*}_{jkR}(\theta_i) - \hat{P}^{*}_{jkF}(\theta_i) \right] \right| \Delta\theta \,+ \\ & \frac{1}{2} \left| \left[u_{j(k+1)} - u_{jk} \right] \left[\hat{P}^{*}_{jkR}(\theta_L) - \hat{P}^{*}_{jkF}(\theta_L) \right] \right| \Delta\theta \,- \\ & \frac{1}{2} \left| \left[u_{j(k+1)} - u_{jk} \right] \left[\hat{P}^{*}_{jkR}(\theta_U) - \hat{P}^{*}_{jkF}(\theta_U) \right] \right| \Delta\theta, \end{aligned} \tag{19}$$

where $\Delta\theta = (\theta_U - \theta_L)/N$.

The Likelihood Ratio Test. The LR test for DIF described by Thissen et al. (1986, 1988, 1993) compares two different models—a compact model and an augmented model. The LR test statistic, G^2, is the difference between the values of -2 times the log likelihood for the compact model ($-2\log L_C$) and -2 times the log likelihood for the augmented model ($-2\log L_A$). The values of the quantity $-2\log L$ can be obtained from the output of the calibration runs from the computer program MULTILOG (Thissen, 1991) and are based on the results over the entire data set following marginal maximum likelihood estimation.

Let y_j be the polytomous score for item j (i.e., $y_j = 1, 2, \ldots, k, \ldots, K_j$) and let

$$u_{jk} = \begin{cases} 1 & \text{if } y_j = k \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

be the indicator variable for item j. Without loss of generality, it can be assumed that all items in the test have the same number of categories K. The category response function describes the probability that $y_j = k$ at ability level θ and is defined as

$$\text{Prob}\left\{y_j = k|\theta, \boldsymbol{\xi}_j\right\} = P_{jk}(\theta) = \prod_{k=1}^{K} P_{jk}(\theta)^{u_{jk}}, \tag{21}$$

where $\boldsymbol{\xi}_j$ represents the vector of item parameters. Under the assumption of local independence, the conditional probability, given θ, of a particular response vector or lth response pattern $\mathbf{y}_l = (y_1, y_2, \ldots, y_J)$ can be written as

$$P(\mathbf{y}_l|\theta) = \prod_{j=1}^{J} \prod_{k=1}^{K} P_{jk}(\theta)^{u_{jk}}, \tag{22}$$

where J is the total number of items in the test. The marginalized probability of a response pattern $\mathbf{y}_l = (y_1, y_2, \ldots, y_J)$ can be written as

$$P(\mathbf{y}_l) = \int P(\mathbf{y}_l|\theta)\pi(\theta|\boldsymbol{\tau})d\theta, \tag{23}$$

where $\pi(\theta|\boldsymbol{\tau})$ is the ability distribution and $\boldsymbol{\tau}$ are the population ability parameters (see Bock & Aitkin, 1981; Thissen et al., 1986). The distribution of ability in the usual IRT model is Gaussian, and, hence, $\boldsymbol{\tau}$ contains μ and σ^2..

To obtain the marginal likelihood, the item response data are summarized to yield raw counts of the number of examinees giving each particular response pattern across all items are used. The counts for group g are denoted by $r_g(\mathbf{y}_l)$ and fill the cell of a K^J contingency table of all possible response patterns for each group. The marginalized probability of observing an examinee in group g with a response pattern $\mathbf{y}_l$ is

$$P_g(\mathbf{y}_l) = \int P(\mathbf{y}_l|\theta)\pi(\theta|\boldsymbol{\tau}_g)d\theta. \tag{24}$$

The likelihood for the complete set of K^J tables for all the groups is proportional to

$$\prod_{g=1}^{G} \prod_{l=1}^{K^J} P_g(\mathbf{y}_l)^{r_g(\mathbf{y}_l)}, \tag{25}$$

where G is the number of groups. The marginal maximum likelihood estimates of the parameters of interest can be obtained using the algorithm described in Bock and Aitkin

(1981). Using default options, the computer program MULTILOG yields the location and scale of θ, arbitrarily set by fixing $\mu_R = 0$ and $\sigma_R^2 = 1$ for the reference group. In addition, a default in MULTILOG also imposes the constraint $\sigma_R^2 = \sigma_F^2$. Then,

$$-2\log L = -2\sum_{g=1}^{G}\sum_{l=1}^{K^J} r_g(\mathbf{y}_l)\log\left[\frac{N_g\hat{P}_g(\mathbf{y}_l)}{r_g(\mathbf{y}_l)}\right], \tag{26}$$

with $N_g = \sum_l r_g(\mathbf{y}_l)$ (i.e., the number of examinees in group g) and $\hat{P}_g(\mathbf{y}_l)$ computed from the marginal maximum likelihood estimates of the parameters. [See Bishop, Fienberg, and Holland (1975) for an extensive discussion of the use of the likelihood ratio statistic in the context of model-fitting for contingency tables.]

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 the augmented model, item parameters for all items except the studied item are constrained to be equal in both the reference and focal groups. These constrained items are referred to as the common or anchor set.

The LR test statistic can be written as

$$G^2 = -2\log L_{\mathrm{C}} - (-2\log L_{\mathrm{A}}) \tag{27}$$

and is distributed as a χ^2 under the null hypothesis with degrees of freedom equal to the difference in the number of parameters estimated in the compact and augmented models (Rao, 1973). When a graded response item with four categories is tested, G^2 is distributed as a χ^2 with four degrees of freedom.

Observed Score Methods for DIF Detection

Two extensions of the Mantel-Haenszel test for dichotomous models (Mantel & Haenszel, 1959) have been proposed by Zwick et al. (1993) for graded response items; the Mantel (1963) test and the generalized Mantel-Haenszel test (Mantel & Haenszel, 1959). The Mantel test assumes that item responses are ordered, whereas the generalized Mantel-Haenszel test assumes that item responses are nominal. The assumption underlying the Mantel test would appear to be theoretically more consistent with the ordered nature of scores used for graded response items. Chang et al. (1996) have described an extension of the simultaneous item bias test (SIBTEST) of Shealy and Stout (1993) for use with polytomous models.

The Mantel Test. Mantel (1963) proposed a test of conditional independence for the case of K ordered categories (see also Agresti, 1990, pp. 283–284). Application of the method in the DIF context involves assigning ordered index numbers to the response categories and then comparing the item means for examinees of the reference and focal groups who have been matched on a measure of proficiency. Ankenmann et al. (1996), Chang et al. (1996), Welch and Hoover (1993), Welch and Miller (1995), and Zwick et al. (1993) investigated this statistic in their studies of DIF methods for the polytomously scored items.

In a DIF study of an item with K ordered response categories, there will be a separate $2 \times K$ contingency table for each level of the matching variable. The data can be arranged into a full $2 \times K \times L$ contingency table, where L is the number of levels of the matching or stratification variable. The total raw score is often used as the matching variable in the Mantel test. For the lth level of the matching variable, for example, a $2 \times K$ contingency table can be constructed to contain the data as shown in Table 1. The values, $Y_1, \ldots, Y_K$, represent the scores that can be obtained on the item. The item scores are typically, but not necessarily, the natural numbers (i.e., $1, \ldots, K$). The values of A_{kl} and B_{kl} denote the number of focal and reference group examinees, respectively, who are at the lth level of the matching variable and received an item score of Y_k. The marginal total of the focal group of the lth level is denoted as N_{Fl}, and that of the reference group as N_{Rl}. The total number of focal and reference group members with an item score Y_k at the lth level of the matching variable is denoted by M_{kl}. The total number of examinees at the lth level of the matching variable is denoted by T_l.

Insert Table 1 about here

Given the marginal totals in each level of the matching variable, under the assumption of conditional independence of the item score variable Y and the group membership variable, the observed sum of the weighted scores

$$\sum_{k=1}^{K} A_{kl} Y_k \tag{28}$$

has its expectation and variance defined as

$$E\left(\sum_{k=1}^{K} A_{kl} Y_k\right) = \frac{N_{Fl} \sum_{k=1}^{K} M_{kl} Y_k}{T_l} \tag{29}$$

and

$$\mathrm{Var}\left(\sum_{k=1}^{K} A_{kl}Y_k\right) = \frac{N_{Fl}N_{Rl}}{T_l^2(T_l-1)}\left[T_l\sum_{k=1}^{K} M_{kl}Y_k^2 - \left(\sum_{k=1}^{K} M_{kl}Y_k\right)^2\right]. \qquad (30)$$

When a dichotomous variable, say X, is used for the group membership variable (e.g., $X_F = 1$ and $X_R = 0$), then the value from the single contingency table is

$$\frac{\left[\sum_{k=1}^{K} A_{kl}Y_k - E\left(\sum_{k=1}^{K} A_{kl}Y_k\right)\right]^2}{\mathrm{Var}\left(\sum_{k=1}^{K} A_{kl}Y_k\right)} \qquad (31)$$

and is the same as the squared point biserial correlation between X and Y, multiplied by the sample size minus one $(T_l - 1)$ for the lth level of the matching variable. Under the null hypothesis of conditional independence, either the point biserial correlation or the value from Equation 31 should be close to zero.

To summarize the association from all L levels of the matching variable, Mantel (1963) proposed the statistic

$$M^2 = \frac{\left[\sum_{l=1}^{L}\sum_{k=1}^{K} A_{kl}Y_k - \sum_{l=1}^{L} E\left(\sum_{k=1}^{K} A_{kl}Y_k\right)\right]^2}{\sum_{l=1}^{L}\mathrm{Var}\left(\sum_{k=1}^{K} A_{kl}Y_k\right)}. \qquad (32)$$

The expected value and the variance are obtained under the assumption of the conditional independence between the item score variable and the group membership variable in each level of the matching variable. Under the null hypothesis of no association, H_0, the test statistic M^2 is distributed as a chi-square with one degree of freedom provided that the total sample size is large. For dichotomous items, this test statistic is identical to the Mantel-Haenszel (1959) statistic without the continuity correction. In DIF applications, rejection of H_0 indicates that examinees in the focal and reference groups who are similar in overall proficiency with respect to the matching variable tend to differ in their average performance on the studied item.

The Generalized Mantel-Haenszel Test. Mantel and Haenszel (1959) described a generalized extension of the ordinary Mantel-Haenszel statistic to the case of $K > 2$ response categories [see also Agresti (1990, pp. 234–235) and Somes (1986)]. The generalized statistic tests the conditional independence for an unordered group variable and K response categories. Application of the method in the DIF context involves assigning nominal numbers to the response categories and then comparing the vectors of the item responses for examinees of the reference and focal groups who have been matched on a measure of proficiency.

Using the notation in Table 1, assuming fixed marginal totals in each level of the matching variable, the observed vector of the number of examinees for $Y_1, \ldots, Y_{K-1}$ of the focal group is

$$\mathbf{a}_l = (A_{1l}, \ldots, A_{kl}, \ldots, A_{(K-1)l})' \tag{33}$$

which has expectation and variance

$$E(\mathbf{a}_l) = N_{Fl}\mathbf{m}_l / T_l \tag{34}$$

and

$$\mathbf{V}_l = \frac{N_{Fl}N_{Rl}}{T_l^2(T_l - 1)} \left[T_l \mathrm{diag}(\mathbf{m}_l) - \mathbf{m}_l\mathbf{m}_l' \right], \tag{35}$$

where

$$\mathbf{m}_l = (M_{1l}, \ldots, M_{kl}, \ldots, M_{(K-1)l})'. \tag{36}$$

The expected value and the variance are based on the conditional independence of the item score variable and the group membership variable. As noted in Agresti (1990), the value

$$[\mathbf{a}_l - E(\mathbf{a}_l)]' \, \mathbf{V}_l^{-1} [\mathbf{a}_l - E(\mathbf{a}_l)] \tag{37}$$

is the Pearson (1900, 1922) chi-square statistic for testing independence, multiplied by a factor $(T_l - 1)/T_l$.

The generalized Mantel-Haenszel statistic summarizes the association from all L levels of the matching variable and is defined as

$$Q^2 = \left[\sum_{l=1}^{L} \mathbf{a}_l - \sum_{l=1}^{L} E(\mathbf{a}_l) \right]' \left[\sum_{l=1}^{L} \mathbf{V}_l \right]^{-1} \left[\sum_{l=1}^{L} \mathbf{a}_l - \sum_{l=1}^{L} E(\mathbf{a}_l) \right]. \tag{38}$$

If we let $\mathbf{a} = \sum_{l=1}^{L} \mathbf{a}_l$, $\mathbf{e} = \sum_{l=1}^{L} E(\mathbf{a}_l)$, and $\mathbf{V} = \sum_{l=1}^{L} \mathbf{V}_l$, then Q^2 can be written in quadratic form as

$$Q^2 = (\mathbf{a} - \mathbf{e})' \mathbf{V}^{-1} (\mathbf{a} - \mathbf{e}). \tag{39}$$

Under the assumption of conditional independence, the test statistic Q^2 has a large-sample chi-square distribution with $K - 1$ degrees of freedom when two groups are used. In case of dichotomous items, this statistic is identical to the Mantel-Haenszel (1959) statistic without the continuity correction. In DIF applications, rejection of H_0 indicates that examinees of the focal and reference groups who are similar in overall proficiency tend to differ in their performance on the studied item.

The SIBTEST for Polytomous Items. Chang, Masseo, and Roussos (1996) describe an extension of the SIBTEST for dichotomous items (Shealy & Stout, 1993) to polytomous items such as graded response items. The amount of DIF measured by this method is

$$B_0(\theta) = E_R(Y_j|\theta) - E_F(Y_j|\theta), \tag{40}$$

where

$$E(Y_j|\theta) = \sum_{k=1}^{K_j} kP_{jk}(\theta), \tag{41}$$

R and F designate the reference group and the focal group, respectively, and Y_j represents the score that can be obtained on the item. The item scores Y_j are possibly, but not necessarily, the natural numbers (i.e., $1, \ldots, K_j$). If there are the same number of categories in all items, then, without loss of generality, we can write $K = K_j$.

A global index of DIF (Shealy & Stout, 1993) is given by

$$\beta_j = \int B_0(\theta)g_F(\theta)d\theta, \tag{42}$$

where $g_F(\theta)$ is the density of θ in the focal group. This is interpreted as the expected amount of DIF experienced by a randomly selected examinee from the focal group.

Two minor modifications to the original SIBTEST are needed to accommodate polytomous data: (1) replacement of the number of items in the SIBTEST with the maximum test score due to polytomous scoring and (2) modification of the matching test reliability estimates used by Shealy and Stout in their regression correction, substituting with Cronbach's alpha for KR20 (Chang, Masseo, & Roussos, 1996).

The test statistic B_j is defined as

$$B_j = \frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)}, \tag{43}$$

where

$$\hat{\beta}_j = \sum_{l=1}^{L} p_l d_l, \tag{44}$$

$d_l = \bar{Y}_{jRl} - \bar{Y}_{jFl}$ is the group difference in performance on the studied item for the examinees in the lth matching variable, p_l is the proportion of the examineees in the lth matching variable (i.e., $p_l = N_l/N$), and

$$\text{s.e.}(\hat{\beta}_j) = \sqrt{\sum_{l=1}^{L} p_l^2 \left[\frac{\text{Var}_{Rl}(Y_j)}{N_{Rl}} + \frac{\text{Var}_{Fl}(Y_j)}{N_{Fl}} \right]}, \qquad (45)$$

where $\text{Var}_{Rl}(Y_j)$ and $\text{Var}_{Fl}(Y_j)$ are the sample variances of the studied item scores for the lth matching variable for examinees in the reference and focal groups, respectively. It can be seen that $N_l = N_{Rl} + N_{Fl}$ and $N = N_R + N_F$, where $N_R = \sum_l N_{Rl}$ and $N_F = \sum_l N_{Fl}$.

The total score for the matching variable can be obtained as

$$X = \sum_{j=1}^{J} X_j, \qquad (46)$$

where J is the total number of items used in the matching variable and X_j are the jth item scores (e.g., $1, \ldots, K_j$). If we assign 1 to K_j for the jth item scores, then X will be J, $J+1$, $\ldots$, $\sum_j K_j$. In this case, the first level, $l = 1$, corresponds to $X = J$, and the highest level $l = L$ corresponds to $X = \sum_j K_j$.

Linking and Purification

Linking Metrics. As in the case for the dichotomous IRT models, the transformation or linking of the metric of the focal group to the metric of the reference group is required under the graded response model before DIF comparisons are made. Baker (1992) extended the test characteristic curve method for linking (Stocking & Lord, 1983) to the case of the graded response model. Recent evidence (Cohen & Kim, in press) suggests that the test characteristic curve method may be more accurate than the minimum chi-square method or mean and sigma methods.

Linking of metrics is required only when item parameter estimates are obtained separately in both groups. DIF comparisons using the LR test procedure do not need to be preceded by linking as item parameters are estimated simultaneously in both groups. In the LR method described by 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 metric of the augmented model, as well as the metric of the compact model, is dependent upon a set of anchor items that are assumed to be free of DIF. Although likelihoods obtained via simultaneous calibration do not require any linking transformation from one metric to another, comparing a compact model to an augmented model does require two separate calibrations for each comparison, one for the compact model and one for the augmented model in which at least one item is unconstrained in the two groups.

Since the methods based on observed scores are not involved with calibration of item parameters, the Mantel test, the Mantel-Haenszel test, and the SIBTEST do not require linking. All three observed score methods, however, assume that there exists a matching variable. The matching variable provides an observed score metric on which item response patterns are compared.

Scale Purification. The linking required for the chi-square test and the area measures may be seriously affected by the presence of DIF items in the set of items used for calculation of the linear transformation coefficients. The results for the dichotomous IRT models indicate that spurious identification of items as DIF or non-DIF may result in the presence of DIF items on the test (cf. Lautenschlager & Park, 1988; Shepard, Camilli, & Williams, 1984).

Two methods, scale purification (Lord, 1980, p. 220) and iterative linking (Candell & Drasgow, 1988), have been recommended for dealing with this problem for the dichotomous IRT models. Iterative linking can be generalized to polytomous IRT models without any modification. Iterative linking described by Candell and Drasgow (1988) proceeds as follows:

1. Estimate item parameters for the reference and focal groups separately.
2. Place the focal group item parameter estimates onto the scale of the reference group.
3. Calculate DIF indices and stop the process if no DIF items are found.
4. Otherwise, remove the DIF items and recalculate the linking coefficients using only the remaining non-DIF items.
5. Calculate DIF indices for all items (including previously identified DIF items).

Steps 4 and 5 are continued until the same set of DIF items is identified on a subsequent iteration. Note that the iterative linking procedure requires item parameters be calibrated one time only in each group. The iterative linking procedure can be applied to the chi-square test and the area measures.

For the LR test, Thissen, et al. (1988, 1993) indicate the need for excluding DIF items from the set of items used as the internal anchor. The approach recommended by Thissen et al. (1988, 1993) for dichotomous items is to first use the Mantel-Haenszel χ^2 (Holland &

Thayer, 1988) to identify DIF items to be removed from the anchor set. Kim and Cohen (1995) describe an iterative procedure for scale purification with the LR test.

Scale purification for the Mantel test and the generalized Mantel-Haenszel test has not been discussed extensively. Zwick et al. (1993) indicated that the studied item should be included in the matching variable. Once DIF items are identified, however, it is possible to remove them from the analysis and sequentially test the remaining items for presence of DIF.

For the SIBTEST, Stout and Roussos (1996) offer the following scale purification steps:

Step 1. Conduct a DIF analysis over the J items of interest. On the J runs of SIBTEST, each item is evaluated sequentially and the remaining $J-1$ items are used to form the matching variable. If any DIF items are detected, those items form the Step 1 suspect set.

Step 2. Conduct the second DIF analysis using the items that were not included in the Step 1 suspect set. If there are J' of such items, then there will be J' subsequent runs of SIBTEST each with $J'-1$ items forming the matching variable. If any additional DIF items are detected, the flagged items form the Step 2 suspect set.

Step 3. Combine the two sets of suspect items and form the Step 3 suspect set. Test each item sequentially in the Step 3 suspect set, one at a time. The unflagged items from Step 2 are used as the matching variable. All items rejected based on a prespecified nominal alpha level are considered to be the DIF items.

Method

Data

Data from Nasser, Takahashi, and Benson (1997) were reanalyzed for purposes of this study. The data were obtained from participants responses to an Arabic version of Sarason's (1984) Reactions to Test (RTT) scale. The RTT scale consists of 40 Likert-type items with four options. The sample consisted of 421 tenth graders from two Arab high schools in the central district of Israel. There were 226 female students and 195 male students in the sample. The purpose of DIF analyses was to compare the item responses of female and male students. For purposes of this study, female students were treated as the reference group and male students as the focal group.

Parameter Estimation and DIF Detection Procedures

Item parameter estimates for the graded response model were obtained using marginal maximum likelihood estimation via the computer program MULTILOG (Thissen, 1991). For Lord's χ_j^2 and the two area measures, item parameter estimates, a_j and b_{jk} $(k = 1, 2, 3)$, were obtained using marginal maximum likelihood estimation via separate calibration runs of MULTILOG.

The computer program EQUATE 2.0 (Baker, 1993) implements the characteristic curve method of equating and was used to obtain the linear coefficients for linking item parameter estimates obtained in the reference and focal groups. The coefficients, A and B, were then used in the following transformations to place the focal group item parameter estimates, a_{jF} and b_{jkF}, and their estimated variances onto the metric of the reference group:

$$a^*_{jF} = a_{jF}/A, \tag{47}$$

$$b^*_{jkF} = A \times b_{jkF} + B, \tag{48}$$

$$\text{var}(a^*_{jF}) = \text{var}(a_{jF})/A^2, \tag{49}$$

and

$$\text{var}(b^*_{jkF}) = A^2 \times \text{var}(b_{jkF}), \tag{50}$$

where $*$ indicates a transformed value. Iterative linking (Candell & Drasgow, 1988) was used with the chi-square test and the two area measures, $Z(S_j)$ and $Z(U_j)$.

For the LR test (Thissen et al., 1988, 1993), the compact model was obtained by calibration over the combined reference and focal groups via the computer program MULTILOG (Thissen, 1991). MULTILOG permits constraints to be placed on the item parameters for estimation of the compact model. The item parameters for all internal anchor items in the augmented model were similarly constrained, and only the item parameters for the studied item were estimated independently in the reference and focal groups.

The metric used in the likelihood ratio test is based upon the set of items contained in the internal anchor. If DIF items are present in the anchor, erroneous identification of items as DIF or non-DIF could result. In this study, we used a sequential approach to purify the anchor set. All DIF items were removed from subsequent anchor sets until no further DIF items were found.

For the Mantel test and the generalized Mantel-Haenszel test, the same iterative purification procedure as used in the LR test was applied. DIF items were sequentially removed until no DIF items were found. Each time a DIF item was identified, it was removed from the matching variable for subsequent DIF comparisons. For the SIBTEST, the scale purification procedure recommended by Stout and Roussos (1996) was applied. Results from each of the DIF detection methods were compared for the initial and final iterations.

Results

Classical Item Statistics

Summary statistics for the reference and the focal groups are presented in Table 2. Item statistics, means and standard deviations, and correlations between the item score and the item-excluded total score are given in Table 3.

Insert Tables 2 and 3 about here

Results of the Chi-Square Test and the Area Measures

Item parameter estimates and estimated variance terms are reported in Table 4. (MULTILOG does not provide estimates of the item parameter covariance terms.) These estimates were used to calculate the metric transformation coefficients, A and B, required for iterative linking. Results for the chi-square test, $Z(S_j)$, and $Z(U_j)$ are presented in Table 5 for the first and final iterations, respectively.

Insert Tables 4 and 5 about here

On the first iteration, five DIF items were detected using the χ_j^2, three DIF items using $Z(S_j)$, and eight items using $Z(U_j)$. Two iterations were required for χ_j^2, $Z(S_j)$, and $Z(U_j)$, respectively. The final iteration yielded the same set of DIF items for the χ_j^2 and $Z(S_j)$ methods and one additional item (item 26) for the $Z(U_j)$ method.

Results of the Tests Based on Observed Scores

Results of the Mantel test, the generalized Mantel-Haenszel test, and the SIBTEST are presented in Table 6 for the first and final iterations. Three iterations were required for M_j^2 and four for Q_j^2. The SIBTEST purification process was implemented as described by Stout and Roussos (1996).

On the first iteration, 11 DIF items were detected using M_j^2 and 10 using Q_j^2. Eight items were identified in the Step 1 set by SIBTEST. The final iterations yielded 11, 14, and 8 DIF items for the M_j^2, Q_j^2, and the SIBTEST, respectively. Although the number of items identified byM_j^2 and the SIBTEST were the same, the actual items were different.

Insert Table 6 about here

Likelihood Ratio Test Results

Results for the analysis of the compact and the augmented models for studying item 1 are given in Table 7. The item parameter estimates and the standard errors for the compact model are given in the two columns to the left of the item numbers. The value of $-2 \log L$ for the compact model was 30250.1 (see footnote at the bottom of Table 7). The item parameter estimates and the standard errors for the augmented model are given to the right of those of the compact model. There are two sets of item parameter estimates for each studied item. The item estimates for the reference and focal groups for item 1 are given in Table 7 to illustrate that there are two sets for each studied item. When item 1 was the studied item, items 2 to 40 were used as the internal anchor set. The value of $-2 \log L$ for the augmented model for item 1 was 30241.7 and is given in the column to the right of the item parameter estimates.

Insert Table 7 about here

For item 1, the likelihood ratio test statistic was $G_j^2 = 30250.1 - 30241.7 = 8.4$. This value was not significant at $\alpha = .01$. Summary results for all 40 items are presented in Table 8. The same 11 items were significant from the first and second (i.e., final) iterations.

Insert Table 8 about here

Comparison of DIF Indices

Similarities between DIF indices was determined by comparing the ranks of the values of one index with the ranks for a second using Spearman's ρ. Values for the two test statistics of the area measures and the SIBTEST statistic B_j were first squared and then ranked. Results were compared for the iterative methods from the first and final iterations only; intermediate results were not included.

Correlations between first and final iterations indicate the impact of the iterative procedures on the magnitude of the DIF indices. Spearman's ρ values for the same DIF index ranged from .877 to 1.000 indicating the iterative procedures had a relatively small impact on the magnitudes of the DIF indices.

There were moderate to strong relationships, among IRT-based DIF indices except for $Z^2(U_j)$. Comparable correlations in the moderate to strong range were observed between the observed score-based indices. Relationships between IRT-based indices and observed score-based indices were also of similar magnitude except for those involving $Z^2(U_j)$.

Insert Tables 9 and 10 about here

Agreement between items identified as functioning differentially by each index was assessed by calculating ϕ coefficients between the sets of detected items. The ϕ coefficients in Table 10 show moderately high to very strong agreement between first and final iterations for the same indices with coefficients ranging from .688 to 1.0. This suggests that the iterative procedures generally had small to no impact on the items identified.

Agreement tended to be moderate to moderately high (.462 to .640) between IRT-based methods. Agreement was modest to moderately high (.288 to .733) between observed score-based methods. Between IRT- and observed score-based methods agreement was generally modest to moderate (ranging from .288 to .657) except for those involving SIBTEST (ranging from .095 to .479.

Discussion

Detection and removal of DIF items on graded response tests is an important concern for test developers. Methods for detection of DIF in this important model are becoming increasingly necessary as performance-type assessments become more widely used. Selection of a DIF detection index, however, is often a difficult and even confusing task. This is especially so when DIF indices do not all identify the same items. In the present paper, several DIF detection indices for graded response items were examined, four IRT-based measures of DIF in the graded response model were described along with three observed score-based DIF measures. DIF detection results for each of these indices using data from a test anxiety scale were then compared.

The DIF detection methods examined all either permitted or required some kind of iterative or sequential removal of DIF items from the test defining the matching variable. The presence of DIF items has been shown to affect the quality of the common metric that is established as well as the quality of the DIF detection in dichotomous IRT models (Kim & Cohen, 1992; Shepard et al., 1984). The use of iterative or sequential methods purification of the test prior to making DIF comparisons, however, did not appear to have reduced differences observed in the DIF items identified by these seven methods. That is, there was strong similarity in the items identified within each method after purification.

There was moderate to high similarity in the magnitudes of six of all DIF indices except for the unsigned area. These results are in general agreement with previous research with these same indices for both dichotomous models and graded response models.

There was also overlap in the set of items identified by each of the seven measures. Unfortunately, the same items were not always identified by each method. This is not an uncommon finding and has led to the usual advice which is to not rely on results from a single DIF detection index. Instead, the recommendation is to use multiple DIF indices. Given the incongruity of agreement among the DIF indices in Table 10, this suggestion seems plausible. In fact, it might make some sense to select DIF detection indices which test for DIF in markedly different ways. In this way, one could hope for some sort of optimal coverage in identifying DIF items.

To some extent, differences in the seven DIF indices can be ascribed, at least in part, to differences in the ways they each identify an item as functioning differentially. Inspection of each of the indices shows that the four IRT-based DIF indices each test for DIF in a different way from one another. Recall that, for the graded response IRT model, DIF was defined as occurring when $T_{jR}(\theta) \neq T_{jF}(\theta)$. The χ_j^2 and G_j^2, however, both test for DIF by examining whether $\xi_{jR} = \xi_{jF}$, that is, whether the item parameters are equal in the reference and focal groups. The signed area measure, S_j, tests DIF as $\int_{-\infty}^{\infty} [T_{jR}(\theta) - T_{jF}(\theta)]\, d\theta$, and the unsigned area measure, U_j, tests this definition in a slightly different way, $\int_{-\infty}^{\infty} \mid T_{jR}(\theta) - T_{jF}(\theta) \mid d\theta$. Both of these approaches are different than the χ_j^2 and G_j^2 and both differ from one another as well. Further, if the distribution of U_j is not normal, the resulting DIF may be tested with an incorrect error term.

A similar point can be made about the DIF that is tested for in the three observed score-based methods, M_j^2, Q_j^2, and B_j. M_j^2 assumes an ordered set of categories, but Q_j^2 assumes nominal categories. When M_j^2 is significant, then the assumption of conditional independence of the item score and the matching variable is rejected. That is, individuals in the focal and reference groups with the same level on the matching variable are likely to differ in their average performance on the studied item. When Q_j^2 is significant, then individuals with the same matching variable but in different groups tend to have different response patterns on the studied item. These two indices differ, in other words, in the way the identify DIF in an item. B_j is a measure of the difference in conditional probabilities of responding the same.

One problem that appears to intrude on the equality of χ_j^2 and G_j^2 results is that these two indices are only equivalent asymptotically. Asymptotic results are not usually obtained in smaller samples or with shorter tests. In addition, estimation errors are present in the variances used to calculate χ_j^2. Further, computer programs such as MULTILOG do not provide the covariances needed for χ_j^2 for the graded response model.

One factor mitigating against use of G_j^2 either with or without iterative purification is that the LR test with iterative purification is far more labor intensive than χ_j^2, and the area measures. Iterative linking methods for χ_j^2, $Z(S_j)$, and $Z(U_j)$ require only a single calibration of item parameters in each group followed by a series of relinking and recalculation of DIF indices. The observed score-based methods, however, were simplest of all to use. The M_j^2 and Q_j^2 do not specifically require purification of the matching variable but it is recommended, and the SIBTEST does have a sequential procedure.

The data presented in this study provide some evidence of the relationships and agreement among these methods. Given the importance of polytomous models such as the graded response model, further empirical evidence would be helpful in assisting test developers to select DIF detection indices. Results of this study can provide useful information about the relationships to expect between various DIF detection methods.

References

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1996, April). *An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87–96.

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

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: The MIT Press.

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

Burden, R. L., & Faires, J. D. (1985). *Numerical analysis* (3rd ed.). Boston, MA: PWS Publishers.

Chang, H.-H., Mazzeo, J., Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33,* 333–353.

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. (in press). A comparison of equating methods under the graded response model. *Applied Psychological Measurement.*

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17,* 335–350.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1995, April). *A Monte Carlo assessment of DFIT with polytomously scored unidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Hogg, R. V., & Craig, A. T. (1978). *Introduction to mathematical statistics* (4th ed.). New York: Macmillan.

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

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. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test in detection of differential item functioning. *Applied Measurement in Education*, *8*, 291–312.

Kim, S.-H. & Cohen, A. S. (in press). An investigation of the likelihood ratio test for detection of differential item functioning under the graded response model. *Applied Psychological Measurement.*

Kim, S.-H., Cohen, A. S., & Baker, F. B. (1996, June). *A comparison of a chi-square test and area measures for the detection of differential item functioning under the graded response model.* Paper presented at the annual meeting of the Psychometric Society, Banff, Alberta, Canada.

Lautenschlager, G. J., & Park, D.-G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement*, *12*, 365-376.

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

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690–700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719–748.

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

Nasser, F., Takahashi, T., & Benson, J. (1997). The structure of test anxiety in Israeli-Arab high school students: An application of confirmatory factor analysis with miniscales. *Anxiety, Stress, and Coping, 10,* 129–151.

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.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables in such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 5, 50,* 157–175.

Pearson, K. (1922). On the χ^2 test of goodness of fit. *Biometrika, 14,* 186–191.

Pearson, K. (1926). On the coefficient of racial likeness. *Biometrika*, *18*, 105-117.

Pine, S. M. (1977). Application of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Application 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., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19,* 353–368.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometric Monographs, 17.*

Samejima, F. (1972). A general model for free response data. *Psychometrika Monograph Supplement*, No. 18.

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46,* 929–938.

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

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

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician, 40,* 106–108.

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

Stout, W., & Roussos, L. (1996). *SIBTEST manual.* University of Illinois at Urbana-Champaign, Department of Statistics, Statistical Laboratory for Educational and Psychological Measurement.

Thissen, D. (1991). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory* (Version 6.0) [Computer program]. Chicago, IL: 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: Lawrence Erlbaum Associates.

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

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6,* 1–19.

Welch, C., & Miller, T. R. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement, 32,* 163–178.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Table 1

Data for the lth Level of the Matching Variable

	Item Score					
Group	Y_1	$\cdots$	Y_k	$\cdots$	Y_K	Total
Focal	A_{1l}	$\cdots$	A_{kl}	$\cdots$	A_{Kl}	N_{Fl}
Reference	B_{1l}	$\cdots$	B_{kl}	$\cdots$	B_{Kl}	N_{Rl}
Total	M_{1l}	$\cdots$	M_{kl}	$\cdots$	M_{Kl}	T_l

Table 2

Summary Statistics for Reference (Female) and Focal (Male) Groups

	Group		
Statistic	Reference	Focal	Total
No. of Subjects	226	195	421
No. of Items	40	40	40
Mean	84.33	72.11	78.67
SD	20.66	17.62	20.23
Coefficient Alpha	.93	.93	.94
SEM	5.35	4.80	5.15

Table 3
Item Statistics for Reference and Focal Groups

	Reference			Focal			Total		
Item	Mean	SD	Corr.	Mean	SD	Corr.	Mean	SD	Corr.
1	2.50	0.93	.52	1.99	0.74	.53	2.26	0.88	.56
2	2.32	1.10	.41	2.03	1.01	.40	2.19	1.07	.42
3	1.43	0.84	.37	1.48	0.76	.37	1.45	0.80	.34
4	1.66	0.92	.50	1.55	0.81	.49	1.61	0.87	.49
5	1.92	0.96	.40	1.52	0.75	.39	1.73	0.89	.43
6	2.36	1.12	.45	1.86	0.96	.56	2.13	1.07	.53
7	1.54	0.85	.44	1.56	0.81	.38	1.55	0.83	.39
8	2.94	1.01	.46	2.54	0.99	.43	2.76	1.02	.48
9	2.01	1.12	.44	1.82	0.93	.26	1.92	1.04	.38
10	1.20	0.60	.18	1.33	0.76	.37	1.26	0.68	.22
11	2.97	1.04	.60	2.21	0.92	.41	2.62	1.06	.58
12	1.87	0.95	.44	1.69	0.79	.43	1.79	0.88	.44
13	2.66	1.05	.20	2.41	1.11	.20	2.54	1.08	.22
14	1.55	0.86	.44	1.34	0.73	.36	1.45	0.81	.43
15	2.60	1.01	.57	2.13	0.87	.60	2.38	0.98	.61
16	2.81	1.05	.58	2.22	0.92	.54	2.53	1.03	.60
17	1.60	0.93	.40	1.60	0.87	.35	1.60	0.90	.36
18	1.38	0.76	.46	1.27	0.64	.46	1.33	0.71	.46
19	2.04	1.02	.56	1.62	0.81	.43	1.85	0.95	.54
20	2.77	0.99	.52	2.37	1.01	.55	2.58	1.02	.56
21	2.47	0.90	.56	2.32	0.93	.70	2.40	0.91	.61
22	2.35	1.09	.41	1.90	1.01	.51	2.14	1.08	.48
23	1.80	0.95	.57	1.36	0.76	.54	1.60	0.89	.59
24	1.54	0.87	.52	1.57	0.81	.56	1.56	0.84	.50
25	1.75	1.02	.44	1.36	0.74	.55	1.57	0.92	.51
26	2.71	1.08	.62	2.00	0.96	.60	2.38	1.08	.65
27	2.54	1.02	.63	2.13	0.94	.54	2.35	1.00	.62
28	1.61	0.95	.45	1.52	0.84	.50	1.57	0.90	.46
29	1.77	0.91	.54	1.69	0.90	.60	1.73	0.90	.55
30	1.85	1.05	.40	1.50	0.78	.46	1.69	0.95	.45
31	2.10	1.10	.53	1.78	0.96	.46	1.95	1.05	.52
32	1.38	0.80	.44	1.32	0.70	.52	1.35	0.75	.46
33	2.33	1.08	.65	1.91	0.94	.60	2.14	1.04	.65
34	2.34	1.01	.65	1.96	0.90	.47	2.17	0.98	.60
35	2.58	1.01	.61	2.10	0.79	.45	2.35	0.95	.58
36	2.31	1.06	.55	1.81	0.94	.49	2.08	1.03	.56
37	2.51	1.07	.54	1.76	0.89	.49	2.16	1.06	.57
38	1.56	0.84	.43	1.49	0.76	.47	1.53	0.81	.43
39	1.98	1.05	.51	1.86	0.84	.30	1.93	0.96	.43
40	2.73	1.08	.60	2.23	0.98	.55	2.50	1.07	.61

Table 4

Item Parameter Estimates for Reference and Focal Groups from Separate Calibration Runs

	Reference				Focal			
Item	$a_{j\mathrm{R}}$(s.e.)	$b_{1j\mathrm{R}}$(s.e.)	$b_{2j\mathrm{R}}$(s.e.)	$b_{3j\mathrm{R}}$(s.e.)	$a_{j\mathrm{F}}$(s.e.)	$b_{1j\mathrm{F}}$(s.e.)	$b_{2j\mathrm{F}}$(s.e.)	$b_{3j\mathrm{F}}$(s.e.)
1	1.44(0.26)	−1.43(0.25)	1.07(0.25)	2.35(0.49)	1.34(0.22)	−2.20(0.37)	0.21(0.17)	1.21(0.24)
2	0.84(0.23)	−1.07(0.32)	1.00(0.40)	2.32(0.71)	0.92(0.19)	−1.31(0.31)	0.44(0.24)	1.52(0.42)
3	0.87(0.28)	0.48(0.30)	2.62(0.83)	3.85(1.26)	0.70(0.21)	1.49(0.56)	2.93(0.90)	4.27(1.26)
4	1.26(0.30)	0.15(0.19)	1.68(0.43)	2.72(0.62)	1.04(0.21)	0.13(0.19)	1.64(0.35)	2.72(0.55)
5	0.95(0.24)	0.23(0.24)	2.29(0.63)	3.94(1.15)	0.81(0.20)	−0.70(0.28)	1.46(0.42)	2.97(0.74)
6	1.45(0.28)	−0.51(0.16)	0.80(0.22)	1.80(0.37)	1.03(0.19)	−1.35(0.29)	0.18(0.21)	1.25(0.30)
7	0.90(0.27)	0.19(0.26)	2.39(0.73)	3.27(0.99)	0.88(0.19)	0.50(0.25)	2.44(0.55)	3.26(0.71)
8	1.05(0.21)	−2.30(0.44)	−0.25(0.21)	1.17(0.33)	1.11(0.21)	−2.67(0.51)	−0.82(0.21)	0.35(0.20)
9	0.60(0.21)	−0.61(0.37)	2.27(0.89)	3.92(1.49)	0.91(0.19)	−0.40(0.22)	0.93(0.31)	1.93(0.48)
10	1.07(0.29)	1.26(0.41)	2.07(0.61)	3.23(0.90)	0.46(0.44)	4.26(2.25)	5.97(4.46)	8.77(4.38)
11	1.09(0.22)	−1.69(0.34)	0.59(0.24)	1.86(0.45)	1.77(0.28)	−2.00(0.28)	−0.67(0.14)	0.08(0.14)
12	1.12(0.25)	−0.40(0.20)	1.56(0.41)	3.24(0.81)	0.96(0.20)	−0.46(0.22)	1.23(0.32)	2.84(0.65)
13	0.48(0.18)	−2.55(1.01)	0.10(0.45)	2.45(1.11)	0.38(0.17)	−4.58(2.97)	−0.79(0.59)	2.44(1.23)
14	1.06(0.32)	1.06(0.38)	2.43(0.73)	3.25(0.98)	0.97(0.23)	0.52(0.24)	2.02(0.48)	3.22(0.75)
15	1.70(0.29)	−1.38(0.20)	0.65(0.19)	1.49(0.29)	1.50(0.22)	−1.87(0.26)	−0.11(0.15)	0.82(0.19)
16	1.40(0.27)	−1.56(0.26)	0.48(0.19)	1.46(0.31)	1.61(0.24)	−1.88(0.26)	−0.47(0.14)	0.38(0.15)
17	0.86(0.27)	0.26(0.28)	1.96(0.68)	3.49(1.24)	0.93(0.21)	0.56(0.25)	1.83(0.48)	3.01(0.74)
18	1.36(0.35)	1.04(0.31)	2.16(0.54)	3.15(0.72)	1.14(0.25)	1.03(0.28)	2.33(0.51)	3.20(0.72)
19	1.08(0.26)	−0.13(0.20)	1.88(0.48)	2.93(0.77)	1.32(0.23)	−0.74(0.18)	0.78(0.22)	1.75(0.31)
20	1.44(0.26)	−1.51(0.24)	0.09(0.17)	1.05(0.26)	1.28(0.21)	−2.36(0.37)	−0.46(0.17)	0.72(0.20)
21	2.08(0.33)	−1.45(0.17)	0.13(0.14)	1.08(0.20)	1.39(0.21)	−2.03(0.31)	0.08(0.16)	1.42(0.26)
22	1.32(0.26)	−0.51(0.18)	0.81(0.25)	1.67(0.38)	0.88(0.19)	−1.52(0.36)	0.33(0.25)	1.55(0.41)
23	1.59(0.33)	0.71(0.22)	1.72(0.36)	2.24(0.46)	1.55(0.22)	−0.22(0.14)	0.91(0.18)	2.05(0.32)
24	1.55(0.31)	−0.03(0.15)	1.55(0.34)	2.18(0.44)	1.12(0.25)	0.50(0.21)	1.80(0.39)	2.82(0.58)
25	1.71(0.33)	0.65(0.20)	1.65(0.33)	2.35(0.49)	1.01(0.21)	0.07(0.19)	1.33(0.32)	2.27(0.48)
26	1.69(0.29)	−0.89(0.16)	0.63(0.19)	1.36(0.28)	1.90(0.26)	−1.73(0.23)	−0.18(0.13)	0.36(0.14)
27	1.50(0.24)	−1.17(0.20)	0.39(0.19)	1.59(0.32)	1.70(0.23)	−1.54(0.20)	−0.18(0.13)	0.93(0.18)
28	1.41(0.31)	0.33(0.18)	1.33(0.33)	2.53(0.55)	0.94(0.22)	0.50(0.23)	1.78(0.44)	2.81(0.65)
29	1.81(0.26)	−0.16(0.13)	0.91(0.19)	1.87(0.33)	1.22(0.22)	−0.24(0.17)	1.25(0.26)	2.57(0.46)
30	1.17(0.28)	0.24(0.21)	2.25(0.54)	2.68(0.66)	0.81(0.18)	−0.15(0.24)	1.35(0.38)	2.50(0.59)
31	1.08(0.23)	−0.34(0.20)	1.27(0.34)	2.16(0.49)	1.19(0.20)	−0.71(0.20)	0.67(0.21)	1.43(0.30)
32	1.69(0.36)	0.77(0.20)	1.76(0.36)	2.43(0.51)	1.08(0.24)	1.26(0.30)	2.03(0.43)	3.24(0.70)
33	1.86(0.31)	−0.63(0.14)	0.68(0.17)	1.52(0.30)	1.86(0.24)	−1.04(0.15)	0.20(0.13)	0.95(0.16)
34	1.18(0.23)	−1.01(0.23)	0.90(0.27)	2.20(0.48)	1.79(0.24)	−1.24(0.17)	0.21(0.13)	1.12(0.19)
35	1.14(0.26)	−1.76(0.35)	0.93(0.28)	2.53(0.60)	1.73(0.24)	−1.68(0.22)	−0.11(0.13)	0.77(0.17)
36	1.20(0.24)	−0.40(0.18)	0.95(0.28)	2.34(0.55)	1.32(0.22)	−1.14(0.21)	0.23(0.16)	1.31(0.26)
37	1.16(0.26)	−0.46(0.19)	1.40(0.36)	2.32(0.54)	1.38(0.23)	−1.57(0.25)	0.02(0.16)	0.89(0.20)
38	1.22(0.26)	0.28(0.19)	1.71(0.40)	3.19(0.75)	0.78(0.19)	0.50(0.29)	2.48(0.62)	4.00(1.01)
39	0.71(0.20)	−1.14(0.39)	1.94(0.63)	4.07(1.27)	1.17(0.20)	−0.44(0.19)	0.74(0.22)	1.93(0.38)
40	1.55(0.27)	−1.28(0.21)	0.29(0.17)	1.29(0.26)	1.69(0.25)	−1.71(0.23)	−0.32(0.14)	0.38(0.15)

Table 5

Lord's χ_j^2, $Z(S_j)$, and $Z(U_j)$ from the First and Final Iterations

	Lord's χ_j^2		$Z(S_j)$		$Z(U_j)$	
Item	First	Final	First	Final	First	Final
1	3.78	3.46	−1.68	−1.63	1.35	1.47
2	1.29	1.21	−0.19	−0.14	0.07	0.23
3	5.76	5.98	1.16	1.22	−0.06	−0.12
4	3.69	4.03	1.04	1.13	0.14	−0.03
5	1.94	1.79	−1.00	−0.95	0.03	0.12
6	1.90	1.84	−0.89	−0.82	0.80	0.76
7	4.95	5.24	0.86	0.93	−0.01	−0.12
8	1.44	1.29	−0.06	−0.06	0.21	0.41
9	6.86	6.78	−1.00	−0.96	0.71	0.80
10	5.81	5.92	1.99	2.01	−0.24	−0.23
11	20.27*	19.65*	−2.44	−2.42	4.95*	5.12*
12	2.74	2.87	0.33	0.40	−0.82	−0.82
13	0.49	0.50	−0.36	−0.36	−0.91	−0.92
14	0.08	0.07	0.01	0.08	−1.20	−1.16
15	2.07	1.83	−0.71	−0.66	0.13	0.40
16	8.51	7.99	−1.35	−1.32	2.67*	2.91*
17	4.46	4.67	0.43	0.49	−0.32	−0.33
18	1.54	1.77	0.89	0.98	−0.13	−0.29
19	3.75	3.47	−1.57	−1.50	1.24	1.43
20	0.30	0.36	−0.18	−0.15	−0.56	−0.77
21	11.74	12.55	2.43	2.47	3.30*	3.08*
22	2.54	2.62	−0.24	−0.18	1.30	1.16
23	3.91	3.60	−0.96	−0.86	0.24	0.50
24	19.70*	20.54*	2.79*	2.88*	2.78*	2.60*
25	2.60	2.71	0.22	0.32	0.85	0.74
26	6.52	6.00	−2.30	−2.24	2.57	2.76*
27	2.37	2.15	−0.15	−0.10	1.03	1.28
28	8.65	9.22	1.94	2.03	1.39	1.23
29	15.23*	16.20*	3.28*	3.38*	3.82*	3.59*
30	1.52	1.54	−0.28	−0.20	−0.25	−0.23
31	1.46	1.34	−0.48	−0.40	0.41	0.63
32	10.75	11.42	2.41	2.51	1.99	1.82
33	1.06	1.01	−0.02	0.07	0.04	0.29
34	8.85	8.47	−0.84	−0.78	2.89*	3.10*
35	14.91*	14.39*	−1.47	−1.44	3.96*	4.13*
36	2.67	2.39	−1.46	−1.40	1.10	1.31
37	12.14	11.65	−3.20*	−3.15*	3.73*	3.87*
38	8.36	8.84	1.90	1.97	0.79	0.70
39	15.05*	14.91*	−0.89	−0.84	1.70	1.80
40	3.58	3.25	−0.76	−0.72	1.50	1.77

*$p < .01$. The critical values are $\chi_4^2 = 13.28$ and $Z = \pm 2.58$.

Table 6

Mantel M_j^2, GMH Q_j^2, and SIBTEST $\hat{\beta}_j$(s.e.) from the First and Final Iterations

Item	M_j^2		Q_j^2		$\hat{\beta}_j$(s.e.)	
	First	Final	First	Final	First	Final
1	6.63*	8.19*	7.23	6.70	−.311(.086)*	−.278(.083)*
2	0.23	0.49	2.20	1.04	−.010(.130)	−.010(.130)
3	14.93*	5.98*	20.40*	20.40*	.227(.070)*	.230(.090)
4	1.64	5.73	1.74	5.13	.123(.090)	.123(.090)
5	3.35	3.61	5.05	3.62	−.209(.092)	−.209(.092)
6	2.13	2.49	7.58	2.73	−.112(.118)	−.112(.118)
7	13.14*	5.24*	14.02*	14.02*	.220(.090)	.220(.090)
8	0.36	0.59	3.40	0.84	−.078(.115)	−.078(.115)
9	0.02	1.26	5.49	11.40*	.239(.134)	.239(.134)
10	7.47*	5.92*	7.55	11.97*	.143(.071)	.143(.071)
11	12.80*	19.65*	13.53*	13.53*	−.422(.105)*	−.570(.098)*
12	2.24	0.95	3.81	4.59	.052(.085)	.052(.085)
13	0.00	0.47	6.42	3.26	.002(.123)	.002(.123)
14	0.13	0.02	0.97	2.52	−.006(.084)	−.006(.084)
15	1.42	0.52	3.12	2.99	−.097(.110)	−.097(.110)
16	7.53*	7.99*	9.84	5.91	−.360(.111)*	−.377(.095)*
17	5.97	6.43	7.37	12.15*	.178(.101)	.178(.101)
18	1.08	1.03	1.74	2.91	−.055(.062)	−.055(.062)
19	1.11	1.08	1.24	0.43	−.013(.102)	−.013(.102)
20	0.22	0.06	0.88	0.84	−.066(.107)	−.066(.107)
21	10.88*	12.55*	12.55*	12.55*	.157(.094)	.157(.094)
22	0.54	2.68	1.97	3.24	−.248(.117)	−.248(.117)
23	6.03	4.47	17.13*	17.13*	−.217(.086)	−.292(.077)*
24	13.67*	20.54*	17.16*	17.16*	.208(.081)	.247(.080)*
25	0.84	1.62	2.63	2.18	−.205(.086)	−.205(.086)
26	12.03*	6.00*	13.64*	13.64*	−.424(.100)*	−.479(.097)*
27	0.67	0.12	1.98	0.43	−.081(.113)	−.081(.113)
28	3.51	2.64	4.92	8.18	.104(.085)	.104(.085)
29	13.34*	16.20*	15.68*	15.68*	.159(.082)	.159(.082)
30	1.60	1.55	4.17	5.45	−.289(.092)*	−.276(.088)*
31	0.33	0.39	0.41	0.94	.095(.115)	.095(.115)
32	5.67	5.94	10.18	18.77*	.069(.074)	.069(.074)
33	0.14	0.82	0.74	2.61	−.040(.105)	−.040(.105)
34	0.49	0.21	0.73	0.31	.030(.107)	.030(.107)
35	2.41	1.02	4.35	6.07	−.012(.103)	−.208(.086)
36	2.71	1.45	4.08	3.36	−.170(.114)	−.170(.114)
37	25.20*	11.65*	25.41*	25.41*	−.503(.109)*	−.556(.099)*
38	5.13	2.61	5.88	6.35	.187(.083)	.187(.083)
39	1.99	2.67	17.13*	17.13*	.306(.094)*	.141(.096)
40	0.57	0.94	3.94	4.01	−.030(.109)	−.030(.109)

*$p < .01$. The critical values are $\chi_1^2 = 6.63$ for M_j^2 and $\chi_3^2 = 11.34$ for Q_j^2.

Table 7

Item Parameter Estimates from the Compact and Augmented Models and the Likelihood Ratio Statistic G_j^2 for Item 1

					Augmented Model										
	Compact Model[a]				Reference/Anchor Item				Focal						
Item	a_j	b_{1j}	b_{2j}	b_{3j}	$a_{j\mathrm{R}}$	$b_{1j\mathrm{R}}$	$b_{2j\mathrm{R}}$	$b_{3j\mathrm{R}}$	$a_{j\mathrm{F}}$	$b_{1j\mathrm{F}}$	$b_{2j\mathrm{F}}$	$b_{3j\mathrm{F}}$	$\hat{\mu}_\mathrm{F}$(s.e.)	$-2\log L$	G_j^2
1	1.56	−1.77	0.41	1.35	1.56	−1.69	0.62	1.81	1.43	−1.98	0.30	1.24	−0.05(.10)	30241.7	8.4
2	0.94	−1.26	0.52	1.63	0.94	−1.26	0.52	1.63							
3	0.64	1.07	3.20	4.73	0.64	1.06	3.18	4.70							
4	1.13	0.04	1.54	2.59	1.14	0.03	1.54	2.59							
5	0.98	−0.34	1.50	2.83	0.98	−0.34	1.50	2.83							
6	1.33	−0.95	0.33	1.24	1.32	−0.96	0.33	1.25							
7	0.79	0.30	2.57	3.51	0.80	0.30	2.56	3.49							
8	1.19	−2.42	−0.64	0.52	1.19	−2.43	−0.65	0.52							
9	0.80	−0.60	1.21	2.36	0.80	−0.60	1.20	2.35							
10	0.56	2.83	4.22	6.39	0.56	2.81	4.18	6.33							
11	1.66	−1.78	−0.31	0.45	1.65	−1.78	−0.31	0.45							
12	1.01	−0.54	1.29	2.90	1.01	−0.54	1.28	2.89							
13	0.48	−3.24	−0.41	2.08	0.48	−3.24	−0.41	2.08							
14	1.08	0.61	1.96	2.94	1.08	0.61	1.96	2.94							
15	1.73	−1.64	0.10	0.91	1.73	−1.64	0.10	0.91							
16	1.69	−1.72	−0.18	0.61	1.68	−1.73	−0.18	0.61							
17	0.81	0.37	1.94	3.38	0.81	0.37	1.93	3.37							
18	1.18	0.97	2.22	3.14	1.18	0.97	2.21	3.13							
19	1.36	−0.56	0.97	1.87	1.36	−0.56	0.97	1.87							
20	1.45	−1.92	−0.29	0.72	1.44	−1.92	−0.29	0.72							
21	1.62	−1.83	0.01	1.16	1.62	−1.83	0.01	1.16							
22	1.16	−1.01	0.43	1.38	1.16	−1.01	0.43	1.38							
23	1.75	0.06	1.02	1.88	1.74	0.06	1.02	1.88							
24	1.10	0.16	1.77	2.69	1.11	0.16	1.76	2.68							
25	1.38	0.24	1.25	2.00	1.38	0.24	1.25	2.00							
26	2.04	−1.32	0.03	0.56	2.03	−1.32	0.03	0.56							
27	1.71	−1.41	−0.05	1.02	1.71	−1.41	−0.05	1.02							
28	1.08	0.33	1.50	2.61	1.08	0.33	1.49	2.60							
29	1.37	−0.30	1.03	2.20	1.38	−0.30	1.02	2.19							
30	1.03	−0.06	1.46	2.25	1.03	−0.06	1.46	2.25							
31	1.22	−0.63	0.76	1.51	1.22	−0.63	0.76	1.51							
32	1.22	0.96	1.87	2.87	1.23	0.96	1.86	2.85							
33	1.98	−0.92	0.28	1.01	1.98	−0.92	0.28	1.01							
34	1.58	−1.20	0.32	1.31	1.59	−1.20	0.32	1.31							
35	1.60	−1.71	0.13	1.09	1.60	−1.71	0.13	1.09							
36	1.39	−0.85	0.38	1.45	1.39	−0.85	0.38	1.45							
37	1.50	−1.05	0.36	1.11	1.50	−1.05	0.36	1.11							
38	0.87	0.34	2.16	3.73	0.88	0.33	2.14	3.71							
39	0.91	−0.84	1.03	2.54	0.92	−0.84	1.03	2.53							
40	1.78	−1.53	−0.17	0.57	1.77	−1.53	−0.17	0.58							

[a] The compact model yielded $-2\log L = 30250.1$.

Table 8

Likelihood Ratio Statistic G_j^2 from the First and Final Iterations

	Iteration	
	First	Final
1	8.4	5.9
2	1.2	1.1
3	22.2*	22.2*
4	6.1	6.2
5	6.0	5.3
6	6.9	5.6
7	9.8	9.7
8	1.5	1.1
9	10.6	10.5
10	17.6*	17.6*
11	28.2*	28.2*
12	6.6	7.1
13	2.8	2.5
14	0.6	0.8
15	5.6	4.2
16	9.3	8.0
17	11.4	12.0
18	2.8	2.9
19	2.9	2.5
20	2.3	2.0
21	17.2*	17.2*
22	5.7	4.2
23	10.8	9.8
24	29.8*	29.8*
25	10.4	8.2
26	11.3	9.2
27	0.4	0.3
28	10.4	11.8
29	15.5*	15.5*
30	16.0*	16.0*
31	0.8	0.7
32	16.4*	16.4*
33	0.9	1.2
34	4.3	2.4
35	15.7*	15.7*
36	3.5	3.0
37	21.3*	21.3*
38	7.6	8.2
39	26.5*	26.5*
40	4.0	3.1

*$p < .01$ with $\chi^2_4 = 13.28$.

Table 9
Spearman's Rho Coefficients Among DIF Indices

		Lord's χ_j^2		$Z^2(S_j)$		$Z^2(U_j)$		G_j^2		M_j^2		Q_j^2		B_j^2	
DIF Index	Iteration	First	Final	First	Final	First	Final	First	Final	First	Final	First	Final	First	Final
Lord's χ_j^2	First														
	Final	.996													
$Z^2(S_j)$	First	.813	.809												
	Final	.813	.810	.994											
$Z^2(U_j)$	First	.647	.634	.526	.518										
	Final	.603	.580	.483	.477	.972									
G_j^2	First	.804	.808	.695	.686	.387	.293								
	Final	.792	.801	.692	.689	.345	.249	1.000							
M_j^2	First	.718	.713	.781	.785	.298	.250	.768	.761						
	Final	.707	.713	.768	.762	.325	.244	.808	.797	.877					
Q_j^2	First	.665	.662	.658	.640	.288	.262	.821	.813	.813	.790				
	Final	.728	.731	.698	.700	.268	.214	.870	.890	.798	.815	.896			
B_j^2	First	.517	.511	.514	.499	.147	.092	.709	.681	.720	.795	.700	.687		
	Final	.562	.554	.566	.551	.213	.156	.735	.708	.757	.779	.680	.683	.904	

Table 10

Phi Coefficients for Agreement Among DIF Indices

		Lord's χ_j^2		$Z(S_j)$		$Z(U_j)$		G_j^2		M_j^2		Q_j^2		B_j	
DIF Index	Iteration	First	Final	First	Final	First	Final	First	Final	First	Final	First	Final	First	Final
Lord's χ_j^2	First														
	Final	1.000													
$Z(S_j)$	First	.640	.640												
	Final	.640	.640	1.000											
$Z(U_j)$	First	.569	.569	.569	.569										
	Final	.528	.528	.528	.528	.928									
G_j^2	First	.462	.462	.462	.462	.532	.473								
	Final	.462	.462	.462	.462	.532	.473	1.000							
M_j^2	First	.493	.493	.493	.493	.577	.657	.550	.550						
	Final	.462	.462	.462	.462	.532	.607	.498	.498	.937					
Q_j^2	First	.493	.493	.493	.493	.433	.518	.550	.550	.733	.679				
	Final	.388	.388	.388	.388	.288	.358	.605	.605	.666	.605	.787			
B_j	First	.095	.095	.095	.095	.219	.329	.392	.392	.433	.532	.433	.288		
	Final	.332	.332	.332	.332	.375	.479	.252	.252	.433	.532	.433	.288	.688	

Acknowledgments

The authors express their gratitude to Fadia Nasser for making her data available for this study, and to Gary J. Lautenschlager and Cynthia A. Searcy for their help in running the computer program for the SIBTEST.

Authors' Addresses

Send correspondence to Allan S. Cohen or James A. Wollack, Testing and Evaluation, University of Wisconsin–Madison, 1025 West Johnson, Madison, WI 53706, Internet: ascohen@facstaff.wisc.edu or jwollack@facstaff.wisc.edu, or Seock-Ho Kim, Department of Educational Psychology, The University of Georgia, 325 Aderhold Hall, Athens, GA 30602, Internet: skim@coe.uga.edu.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM028365

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A Comparison of Item Response Theory and Observed Score DIF Detection measures for the Graded Response Model

Author(s): Allan S. Cohen, Seock-Ho Kim, & James A. Wollack

Corporate Source: University of Wisconsin-Madison NCME

Publication Date: April, 1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
☑	□	□
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproductio n by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Seock Ho Kim

Printed Name/Position/Title: Seock-Ho Kim, Assistant Professor

Organization/Address: The University of Georgia
325 Aderhold Hall
Athens, GA 30602

Telephone: (706) 542-4224

FAX: (706) 542-4240

E-Mail Address: skim@coe.uga.edu

Date: 3/3/98

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

The Catholic University of America
ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
Washington, DC 20064
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com