This paper is concerned with statistical issues in differential item functioning (DIF). Four subsets of large scale performance assessment data from the Georgia Kindergarten Assessment Program-Revised (N=105,731; N=10,000; N=1,000; and N=100) were analyzed using three DIF detection methods for polytomous items to examine the congruence among the DIF detection methods. The DIF detection methods were the likelihood ratio test, the Mantel test, and the generalized Mantel-Haenszel test. Results indicated some agreement among the DIF detection methods within each sample and across the samples except for N=100. Because statistical power is a function of the sample size, however, the DIF detection results from extremely large samples are not useful. As alternatives to the DIF detection methods, four model-based indices of standardized impact and four observed-score indices of standardized impact for polytomous items were obtained and compared for N=105,731. (Contains 3 figures, 10 tables, and 55 references.) (Author/SLD)
An Investigation of the Likelihood Ratio Test, the Mantel Test, and the Generalized Mantel-Haenszel Test of DIF

Seock-Ho Kim
The University of Georgia

April 27, 2000
Running Head: DIF DETECTION AND INDICES OF IMPACT

An Investigation of the Likelihood Ratio Test, the Mantel Test, and the Generalized Mantel-Haenszel Test of DIF

Abstract

This paper is concerned with statistical issues in differential item functioning (DIF). Four subsets of large scale performance assessment data ($N = 105,731$, $N = 10,000$, $N = 1,000$, and $N = 100$) were analyzed using three DIF detection methods for polytomous items to examine the congruence among the DIF detection methods. Results indicated some agreement among the DIF detection methods within each sample and across the samples except for $N = 100$. Because statistical power is a function of the sample size, however, the DIF detection results from extremely large samples are not useful. As alternatives to the DIF detection methods, four model-based indices of standardized impact and four observed-score indices of standardized impact for polytomous items were obtained and compared for $N = 105,731$.

Key words: differential item functioning, generalized Mantel-Haenszel test, graded response model, item response theory, indices of impact, likelihood ratio test, Mantel test.
Introduction

For many years, topics related to item bias, test bias, and unfairness in testing have been the source of many perplexing debates in the educational measurement and educational policy communities (e.g., Berk, 1982; Holland & Wainer, 1993; Wainer & Braun, 1988). In the past, differential item function (DIF) has been referred to as 'item bias' in the literature. DIF is a generic term which indicates that some effort has been made to condition on proficiency or total test scores before examining differences in item performance of subgroups of examinees. For dichotomously scored items, 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 (cf. Pine, 1977).

The presence of DIF items on a test poses a serious threat to fairness in test use and validity of the interpretation of test scores. In this regard, Standard 7.3 in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) describes the following:

When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups (p. 81).

Likewise, one of the guidelines in the Code of Fair Testing Practices in Education (APA, 1988) specifies that:

Test developers should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

In order to make a fair test, test developers should investigate empirically the performance of examinees from different sociocultural backgrounds, and give test users an opportunity to evaluate the extent of the inappropriate characteristics of the test and the differences in test performance. A DIF analysis for a test, hence, can be seen as an essential step to protect the rights of test takers and the general public and as a indispensable tool for test developers to demonstrate the fairness of the test.
Although DIF research for the last several decades has focused primarily on dichotomously scored items and tests, recent efforts to develop alternative measurement methods, such as performance assessment, authentic assessment, and portfolio assessment, have sparked interest in looking at other types of DIF especially in polytomously scored items. It is important to note that there is some emerging evidence that greater discrepancy can be found in performance of ethnic groups under performance assessment (Dunbar, Koretz, & Hoover, 1991; Zwick, Donoghue, & Grima, 1993a), even though there exists a belief that performance assessment is intrinsically more fair than the usual tests with objective (e.g., multiple-choice) formats.

During the 1990's a number of procedures were proposed for detection of DIF in polytomously scored items (e.g., Chang, Mazzeo, & Roussos, 1996; Cohen, Kim, & Baker, 1993; Miller & Spray, 1993; Raju, van der Linden, & Fleer, 1995; Welch & Hoover, 1993; Zwick et al., 1993a). A recent survey of many of these methods was provided by Potenza and Dorans (1995). The focus of this study was on the three DIF detection methods for polytomous items; the likelihood ratio test (Wainer, Sireci, & Thissen, 1991), the Mantel (1963) test, and the generalized Mantel-Haenszel (GMH) test (Mantel & Haenszel, 1959). The likelihood ratio test can be seen as an item response theory (IRT) model based method, whereas the Mantel test and the GMH test are extensions of the Mantel-Haenszel (1959) procedure and can be classified as the observed score methods.

The likelihood ratio test was chosen because the invariance principle of IRT provides an ideal framework for DIF detection. In previous studies the likelihood ratio test has been found to yield a good Type I error control for polytomous items (Kim & Cohen, 1998) and good power for tests which combine both dichotomous and polytomous items (Ankenmann, Witt, & Dunbar, 1999). The Mantel test and the GMH test were chosen because these have been found to yield good Type I error control and power for tests which combine both dichotomous and polytomous items, especially when the ability distributions of the groups compared were similar (Ankenmann et al., 1999; Chang et al., 1996; Welch & Hoover, 1993; Zwick et al., 1993a). Zwick et al. (1993a) reported, however, that the Mantel test and the GMH test were sensitive to different types of DIF. The Mantel test seems to be an effective DIF detection method when the between group difference in item means is of primary interest; and the GMH test might be more useful when the interest is on the entire response distributions of the groups (Zwick et al., 1993a).
The present paper investigated the applicability of the three DIF detection methods to large scale performance assessment data when different sample sizes were employed in the analyses. The next section presents the three DIF detection methods used in this study. Because the graded response model was used in the likelihood ratio test, a formal definition of DIF under the graded response model and the null hypothesis tested in the Mantel test and the GMH test were included. The following section presents the comparisons of the three DIF detection methods based on the DIF analyses of four subsets of the performance assessment test data from the Georgia Kindergarten Assessment Program-Revised. Problems with applying DIF detection methods to large data were illustrated. Next, as alternatives to DIF statistics, descriptive indices that characterize the amount of DIF were presented (see Dorans & Kulick, 1986; Wainer, 1993; Zwick et al., 1993a). The four model-based indices of standardized impact as well as the four observed score indices of standardized impact were presented. The final section contains discussion and suggestions for DIF detection using large test data.

Three DIF Detection Methods

Likelihood Ratio Test

Samejima's graded response model was employed in the likelihood ratio test. 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 $\theta$. For an item with $K_j$ categories, $P_{jk}(\theta)$ is defined as

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

where $\theta = 1, \ldots, K_j$. In Equation 1, $P_{jk}^*(\theta)$ is the boundary response function given by

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

where $\alpha_j$ is the discrimination parameter for item $j$, $\beta_{jk}$ is the location parameter of response category $k$ for item $j$, and $\theta$ 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

$$P_{jk}(\theta) = P_{j(k-1)}^*(\theta) - P_{jk}^*(\theta).$$
DIF under the model based methods is defined in terms of item true score functions. For a polytomously scored item such as a graded response item, 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

\[ TS(\theta) = \sum_{j=1}^{J} \sum_{k=1}^{K_j} y_{jk} p_{jk}(\theta), \]  

(4)

where \( J \) is the number of items in the test and \( y_{jk} \) 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 3 it would be 3. The item true score function for a single item \( j \) can be defined as

\[ T_j(\theta) = \sum_{k=1}^{K_j} y_{jk} p_{jk}(\theta). \]  

(5)

For a dichotomous item under IRT, the IRF for the correct response is the item true score function.

In the typical DIF study, there are two groups of examinees, the reference group and the focal group. 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, item \( j \) is identified as a DIF item, when \( T_{jR}(\theta) \neq T_{jF}(\theta) \). Further, the item true score functions from the reference and focal groups are identical if the boundary 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.

The equality of sets of item parameters for graded response items can be tested using several different approaches. The likelihood ratio test for DIF described by Thissen, Steinberg, and Gerrard (1986) and Thissen, Steinberg, and Wainer (1988, 1993) compares two different models; a compact model, in which the parameters for the same item are constrained to be identical in the two groups, and an augmented model, in which at least one item is not constrained to have equal parameters in the two groups. The likelihood ratio 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 dataset following marginal maximum likelihood estimation.
Let \( y_j \) be the polytomous score for item \( j \) (e.g., \( y_j = 1, \ldots, K_j \)) and let

\[
\begin{align*}
u_{jk} &= \begin{cases} 
1 & \text{if } y_j = k \\
0 & \text{otherwise}
\end{cases} 
\end{align*}
\]

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 \( \theta \), and is defined as

\[
\text{Prob} \{ y_j = k | \theta, \xi_j \} = P_{jk}(\theta) = \prod_{k=1}^{K} P_{jk}(\theta)^{u_{jk}},
\]

where \( \xi_j \) represents the vector of item parameters. Under the assumption of local independence, the conditional probability, given \( \theta \), of a particular response vector or \( l \)th response pattern, \( y_i = (y_1, y_2, \ldots, y_J) \), can be written as

\[
P(y_i | \theta) = \prod_{j=1}^{J} \prod_{k=1}^{K} P_{jk}(\theta)^{u_{jk}},
\]

where \( J \) is the total number of items in the test. The marginalized probability of response pattern \( y_i = (y_1, y_2, \ldots, y_J) \) can be written as

\[
P(y_i) = \int P(y_i | \theta) g(\theta | \tau) d\theta = \int P(y_i | \theta) dG(\theta | \tau),
\]

where \( g(\theta | \tau) \) is the ability distribution and \( \tau \) 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, \( \tau \) contains \( \mu \) and \( \sigma^2 \).

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. The counts for group \( g \) are denoted by \( r_g(y_i) \), 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 response pattern \( y_i \) is

\[
P_g(y_i) = \int P(y_i | \theta) g(\theta | \tau_g) d\theta = \int P(y_i | \theta) dG(\theta | \tau_g).
\]

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

\[
\prod_{g=1}^{G} \prod_{l=1}^{K^J} P_g(y_i)^{r_g(y_i)},
\]

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, MULTILOG calibration yields the location and scale of \( \theta \), arbitrarily set by fixing \( \mu_R = 0 \) and \( \sigma_R^2 = 1 \) for the reference group. In addition, a default in MULTILOG imposes the constraint \( \sigma_R^2 = \sigma_P^2 \), while \( \mu_F \) for the focal group is estimated from data. Then,

\[
-2 \log L = -2 \sum_{g=1}^{G} \sum_{i=1}^{K^g} r_g(y_i) \log \left( \frac{N_g \hat{P}_g(y_i)}{r_g(y_i)} \right),
\]

with \( N_g = \sum_i r_g(y_i) \) (i.e., the number of examinees in group \( g \)) and \( \hat{P}_g(y_i) \) 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.]

The likelihood ratio test statistic can be written as

\[
G^2 = -2 \log L_C - ( -2 \log L_A )
\]

and is distributed as a \( \chi^2 \) 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 three categories is tested, \( G^2 \) is distributed as a \( \chi^2 \) with 3 degrees of freedom.

**Mantel Test**

Two extensions of the Mantel-Haenszel test of DIF for dichotomous items (see Holland & Thayer, 1988) have been used in Zwick et al. (1993a) for polytomously scored items; the Mantel test (1963) and the GMH test (Mantel & Haenszel, 1959). The Mantel test assumes that item responses are ordered, whereas the GMH 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 in the graded response items.

The Mantel test is a test of conditional independence for the case of \( K \) ordered categories (see 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. It is customary to use the total or summed scores that include the studied item as the matching variable (Zwick et al., 1993a).

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 variable. For the $l$th 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 values of $A_{kl}$ and $B_{kl}$ denote the number of focal and reference group examinees, respectively, who are at the $l$th level of the matching variable and received an item score of $Y_k$. The marginal total of the focal group of the $l$th 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 item score $Y_k$ at the $l$th level of the matching variable is denoted by $M_{kl}$. The total number of examinees at the $l$th level of the matching variable is denoted by $T_l$.

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 for the focal group,

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

has its expectation and variance defined as

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

and

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

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

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

$$\text{Var} \left( \sum_{k=1}^{K} A_{kl} Y_k \right)$$

and is the same as the squared point biserial correlation between $Y$ and $Z$, multiplied by the sample size minus one $(T_l - 1)$ for the $l$th level of the matching variable. Under the
null hypothesis of conditional independence, either the point biserial correlation or the value from Equation 17 should be close to zero for each level of the matching variable.

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

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

(18)

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.

**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; Somes, 1986). The GMH statistic tests the conditional independence for a group variable and an item with $K$ unordered 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

$$a_l = (A_{l1}, \ldots, A_{lk}, \ldots, A_{lK-1})'$$

(19)

which has expectation

$$E(a_l) = N_{Fl} m_l / T_l$$

(20)
and variance

\[ V_l = \frac{N_F l N_R l}{T_l^2 (T_l - 1)} [T_l \text{diag}(m_l) - m_l m_l'] , \tag{21} \]

where

\[ m_l = (M_{1l}, \ldots, M_{kl}, \ldots, M_{(K-1)l})' . \tag{22} \]

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

\[ [a_l - E(a_l)]' V_l^{-1} [a_l - E(a_l)] \tag{23} \]

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

\[ Q^2 = \left[ \sum_{l=1}^{L} a_l - \sum_{l=1}^{L} E(a_l) \right]' \left[ \sum_{l=1}^{L} V_l \right]^{-1} \left[ \sum_{l=1}^{L} a_l - \sum_{l=1}^{L} E(a_l) \right] . \tag{24} \]

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

\[ Q^2 = (a - e)' V^{-1} (a - e) . \tag{25} \]

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 in the focal and reference groups, who are similar in overall proficiency, tend to differ in their performance on the studied item.

**Analyses of GKAP-R Data**

**Data**

To compare the three DIF detection methods (i.e., the likelihood ratio test, the Mantel test, and the GMH test), the 1998 Fall data of the Baseline version of the Georgia Kindergarten Assessment Program-Revised (GKAP-R) were analyzed. The Baseline version of the GKAP-R is a performance assessment rating instrument that consists of ten polytomously scored
items with three ordered categories. The scores used in the study were 0, 1, and 2. The full description of the GKAP-R can be found in the Georgia Department of Education web site (http://www.doe.k12.ga.us/sla/ret/gkap.html).

A total of 105,731 students who did not have any omitted or unreached responses were used. There were 55,017 male students and 50,714 female students in this sample. Three other samples with equal numbers of male and female students were randomly formed from the 105,731 students to investigate the effect of the sample size on DIF detection; \( N = 10,000, N = 1,000, \) and \( N = 100. \) The purpose of DIF analyses was to compare the item responses of male and female students. Female students were treated as the reference group and male students were treated as the focal group in DIF analyses. The summary statistics from the male students, the female students, and the total group are presented in Table 2 for the four samples. The average scores were higher for the female students than for the male students except for \( N = 100. \)

---

**Preliminary Analyses**

Before beginning the DIF analyses, classical item statistics were obtained for each item from the \( N = 105,731 \) sample. The results are presented in Table 3. For the total group the range of item means was from .80 (item 3) to 1.78 (item 6). The same items also determined the ranges of item means for the male students and the female students. All of the item means from the female students were higher than the respective item means form the male students. For the total group the item and corrected total score correlations were very high, ranging from .42 (item 5) to .66 (item 7). Similar patterns were observed for the male and female students.

---

The likelihood ratio test was performed under the graded response model. Because the graded response model is a unidimensional IRT model, dimensionality of data was examined. A rough procedure is to computer the latent roots of the polychoric item intercorrelation matrix (cf. Lord, 1980). When the first root is large compared to the second and the
second root is not much larger than any of the others, then the items can be seen as approximately unidimensional. The latent roots of the polychoric item correlation matrix from each sample, obtained from the exploratory factor analysis using the computer program LISCOMP (Muthén, 1988), are presented in Table 4. Figure 1 also shows the ten latent roots for the samples of \( N = 105,731, N = 10,000, N = 1,000, \) and \( N = 100. \) The plots suggest that the items are reasonably unidimensional.

For the likelihood ratio test, the compact model was obtained by calibration over the combined reference and focal groups using 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. For the Mantel test and the GMH test, the summed scores that included the studied item were used as the matching variable.

**Results**

Results for the analysis of the compact and the augmented models for studying item 1 for \( N = 105,731 \) are given in Table 5. The item parameter estimates and the standard errors for the compact model are given in the three columns to the right of the item numbers. Note that the estimated standard errors were extremely small due to the large sample size. The value of \( -2 \log L \) for the compact model was 73622.1 (see footnote at the bottom of Table 5). 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 the studied item. The item parameter estimates for item 1 for the reference group and the focal group, respectively, are given in Table 5 to illustrate that there were two sets of estimates for each studied item. When item 1 was the studied item, items 2 to 10 were used as the internal anchor set. The estimated focal group mean ability parameter was \(-.14\) from the augmented model. The value of \( -2 \log L \) for the augmented model with item 1 as the studied item was 73505.2. For item 1, the likelihood ratio test statistic was \( G^2 = 73622.1 - 73505.2 = 116.9. \) This value was statistically significant at \( \alpha = .01. \)
Summary results from the likelihood ratio test for all 10 items for \( N = 105,731 \) are presented in Table 6. The same \(-2 \log L = 73622.1\) for the compact model was used to obtain the likelihood ratio test statistics \( G^2 \) for all items. Table 6 contains item parameter estimates from the reference and focal groups as well as the estimated focal group mean ability parameters.

Results of the likelihood ratio test, the Mantel test and the GMH test are presented in Table 7 for \( N = 105,731 \), \( N = 10,000 \), \( N = 1,000 \), and \( N = 100 \). The sample size seems to determine the number of significant statistics for all three DIF detection methods. When \( N = 105,731 \), all DIF statistics except item 1 for \( M^2 \) were statistically significant, and all but item 1 were identified as DIF items at a nominal alpha level .01. When \( N = 10,000 \), five items (items 5, 7, 8, 9, and 10) for \( G^2 \) and the same six items (items 3, 5, 7, 8, 9, and 10) for \( M^2 \) and for \( Q^2 \) were identified as DIF items at \( \alpha = .01 \). When \( N = 1,000 \), item 10 was the only item detected as a DIF item by all three methods at \( \alpha = .01 \). None of the items, however, were identified as DIF items when \( N = 100 \).

Similarities between DIF detection statistics can be determined by comparing the ranks of the values of one index with the ranks for a second using Spearman's correlation (see Table 8). Correlations within the same sample were very high except for \( N = 100 \). Correlations between two observed score methods, the Mantel test and the GMH test, were higher than other correlations. There were positive relationships among the three DIF detection statistics across different sample sizes except for \( N = 100 \). Note that the agreement among the three DIF detection methods can also be obtained using correlation coefficients of the binary variables based on DIF identification results at \( \alpha = .01 \).
Indices of Standardized Impact

Descriptive DIF Measure

All three methods used in the previous section are primarily aimed at detection of DIF. As for the case of the null hypothesis testing in practice, it is not expected that any two populations (e.g., male students and female students) in DIF analyses have literally the same sets of item parameters or item means. Because statistical power is a function of the sample size (Cohen, 1988, p. 14), a small difference in population parameters would result in a statistically significant difference when we have a large sample. In other words, we would always expect to reject the null hypothesis when the sample size is huge and statistical power is sufficiently great. When \( N = 105,731 \), all GKAP-R items except item 1 for the Mantel test were identified as DIF items by the three DIF detection methods. This might not be an acceptable conclusion.

When the sample size is large, we may use a descriptive measure of DIF called standardized impact as a viable alternative to the DIF detection methods. The standardized impact can be obtained for both model-based procedures (Wainer, 1993) and for empirically based (i.e., observed score) procedures (Dorans & Kulick, 1986; Dorans & Schmitt, 1991; Zwick et al., 1993a). These two types of indices of standardized impact are presented below. At the outset it should be emphasized that in the context of standardized impact we are not, in general, interested in testing the hypothesis of the difference in true score functions or of independence of item performance by gender.

Model-Based Indices

Wainer (1993) provided four indices of standardized impact for dichotomous IRT models. For polytomously scored items, the four indices of standardized impact can be defined as

\[
T(1) = \int_{-\infty}^{\infty} [T_R(\theta) - T_F(\theta)] dG_F(\theta),
\]

\[
T(2) = N_F T(1),
\]

\[
T(3) = \int_{-\infty}^{\infty} [T_R(\theta) - T_F(\theta)]^2 dG_F(\theta),
\]

and

\[
T(4) = N_F T(3),
\]
where $T_R(\theta)$ and $T_F(\theta)$, without subscript $j$, are the true score functions from the reference group and the focal group, respectively, $G_F(\theta)$ is the proficiency distribution for the focal group, and $N_F$ is the total number of examinees in the focal group.

These indices were related to the earlier descriptive measures that assess the amount of DIF by the area between the two item response functions of dichotomous IRT models (e.g., Linn, Levine, Hasting, & Wardrop, 1980; Raju, 1988; Rudner, 1977). According to Wainer (1993), the index of standardized impact, $T(1)$, can be seen as the average impact for each person in the focal group; $T(2)$ is a measure of total impact that may be useful when the measures of impact are obtained for various focal groups; $T(3)$ is the squared standardized impact where the non-uniform type DIF can be captured by the measure; and $T(4)$ is the total squared impact.

Before presenting values of the indices for the GKAP-R items, let us illustrate the calculation or steps of obtaining $T(1)$ using item 10. All plots needed for the calculation of $T(1)$ are presented in Figure 2. The item parameter estimates were from Table 5.

Insert Figure 2 about here

In Figure 2, the top two plots are the category response functions of item 10 for male and for female, respectively. The second row contains the two boundary response functions of item 10 for male and female. The third row contains the respective item true score functions of item 10 for male and female. The fourth row presents two item true score functions and two ability distributions for male (the focal group) and female (the reference group). Note that the proficiency distribution for the focal group was Gaussian with estimated mean $-0.14$ and variance 1, that is, $g_F(\theta) = N(-0.14, 1)$. The fifth row shows the difference in the two item true score functions and the focal group ability distribution. The final plot is the standardized impact obtained from the multiplication of the difference in the item true score functions and the focal group ability distribution, $[T_R(\theta) - T_F(\theta)] g_F(\theta)$. Actually the $\theta$ was not yet integrated in the final plot. The contribution of different proficiency levels can be seen in the impact plot. When integration is performed with regard to $\theta$, then $T(1)$ is obtained. The values of the model-based indices of standardized impact for GKAP-R items are presented in Table 9.

Insert Table 9 about here

15
then the value of $T(1)$ is between $\pm 2$. When we have several items with different scoring, we may use an index such as

$$T(1)/R,$$

(30)

where $R$ is the range of item scores. The possible values of the index will be limited within $\pm 1$. Tentatively, if $|T(1)|$ is greater than .1 (i.e., $|T(1)|/R$ is greater than .05 when item scores are 0, 1, and 2), then we may conclude the item is deemed to require close examination. Justification of these cutoff values are presented below in the context of the observed score indices of standardized impact.

Wainer (1993) presented ways of measuring the variability of the indices of standardized impact. One method was based on multiple imputation (Rubin, 1987) utilizing the duality diagram concept (Ramsay, 1982) that involved the standard errors of the item parameter estimates. Note that the size of the estimated standard errors is certainly dependent upon the number of examinees used in calibration. As the sample size increases, the variability of the indices of standardized impact decreases. Hence, it may be better to use these indices of standardized impact in a descriptive manner when we have a large sample.

### Observed Score Indices

There are two empirical indices that can be considered as descriptive DIF measures (see Zwick et al., 1993a, 1993b); one stemming from the Mantel test (i.e., the standardized mean difference, SMD) and the other supplements the GMH test (i.e., the Yanagawa and Fujii statistic). Only the SMD is related to the model-based index of standardized impact. The SMD was an extension of the descriptive DIF measure for dichotomous items (Dorans & Kulick, 1986) and first presented in Dorans and Schmitt (1991).

The observed score index of standard impact is

$$T'(1) = \sum_{l=0}^{L} [E_R(Y|X = l) - E_F(Y|X = l)] \frac{N_{Fl}}{N_F},$$

(31)

where

$$E_R(Y_j|X = l) = \frac{\sum B_{ki}Y_k}{N_{Rl}}$$

(32)

and

$$E_F(Y_j|X = l) = \frac{\sum A_{ki}Y_k}{N_{Fl}}$$

(33)
are the expected item scores given the summed score $X = l \ (l = 0(1)L)$ for the reference group and the focal group, respectively, and

$$\frac{N_{Fl}}{\sum_{l=0}^{L} N_{Fl}} = \frac{N_{Fl}}{N_F}$$

is the relative frequency of the focal group examinees for level $l$. The above index is defined as $T'(1)$ because it is a counterpart, which is obtained from the observed scores, to the model-based index of standardized impact $T(1)$. This statistic is in fact the same as $-1$ times Dorans and Schmitt’s (1991) standardized p-difference and Zwick et al.’s (1993a) SMD due to the reversal of the reference group and the focal group. The other observed indices of standard impact are

$$T'(2) = N_F T'(1) = \sum_{l=0}^{L} [E_R(Y|X = l) - E_F(Y|X = l)] N_{Fl},$$

$$T'(3) = \sum_{l=0}^{L} [E_R(Y|X = l) - E_F(Y|X = l)]^2 \frac{N_{Fl}}{N_F},$$

and

$$T'(4) = N_F T'(3) = \sum_{l=0}^{L} [E_R(Y|X = l) - E_F(Y|X = l)]^2 N_{Fl}.$$

Let us illustrate the calculation of the observed score index of standardized impact using item 10 from the GKAP-R data. When the summed score was used as a criterion variable instead of $\theta$, we may obtain empirical trace lines of the three categories of item 10 for male and for female, respectively (see Figure 3, the first row). The summary frequencies for item 10 are presented in Table 10. The second row of Figure 3 contains empirical boundary lines for male and female. The expected item scores for male and female are presented in the third row of Figure 3. Two expected item scores and the relative frequency of summed scores are presented in the fourth row of Figure 3. The fifth row of Figure 3 contains the difference in expected item scores and the relative frequency of the focal group. The observed score index of standard impact is presented at the bottom of Figure 3 without having $\Sigma_i$ performed. We are not interested in the statistical significance testing of the SMD, nevertheless the two different variances of the SMD have been presented by Zwick and Thayer (1996). The two variances are presented in the Appendix.
The four observed score indices of standard impact for the GKAP-R items are presented in Table 9. Note that in order to remove the effect due to the range of the item scores, $T'(1)$ can be divided by the range. If we use $T'(1)/R$ instead of $T'(1)$, then the possible values of the index will be limited within $\pm 1$. Since the item scores were 0, 1, and 2, $T'(1)$ can range from $-2$ to $2$. Positive values of $T'(1)$ indicate that the item favors the reference group, while negative values of $T'(1)$ indicate the opposite. Following Dorans and Holland (1993) (see also Dorans & Kulick, 1986; Dorans & Schmitt, 1991), we may consider $T'(1)$ values between $-.10$ and $.10$ (i.e., $|T'(1)|/R = |T'(1)|/2 < .05$) negligible. $T'(1)$ values between $-.20$ and $-.10$ and between $.10$ to $.20$ (i.e., $.05 \leq |T'(1)/2| < .10$) should be inspected to ensure that no possible effect is overlooked. According to Dorans and Kulick (1986), this might include some items that would be deemed acceptable after close examination. $T'(1)$ values outside the $-.20$ to $.20$ range (i.e., $.10 \leq |T'(1)/2|$) are unusual and require very careful examination.

According to the above flagging cutoffs, item 10 seems to require a closer examination. The positive value of $T'(1)$ indicates that the item favors the reference group, female students. Item 10 is the teacher’s rating (0, 1, 2, where 2 indicates positive approval) of whether a student follows the teacher’s directions. Although conditioned upon the summed scores, the female students seem more likely to follow teachers’s directions than the male students. This difference in compliance between female and male preschool students was not unexpected. It is known that girls are more compliant than boys to the requests and demands of parents, teachers, and other authority figures (Shaffer, 2000). Note that $T'(3)$ might be useful when we have items that exhibit non-uniform DIF, and $T'(2)$ and $T'(4)$ might be helpful when we analyze multiple groups.

Because the observed score indices are counterparts to the model-based indices where the model-based latent $\theta$ values were replaced by observed summed scores $X$, we may apply the same flagging cutoffs to the model-based indices. Hence, we should examine the item with care when we find more than five percent difference (based on the range of item scores) in the indices of standardized impact (i.e., $T(1)$ and $T'(1)$).

**Discussion**

Detection and removal of DIF items on tests with polytomously scored items has become an important concern for both test developers and measurement specialists. Selection of a DIF
detection method, however, is often a difficult and even confusing task. This is especially so when DIF detection methods do not all identify the same items because each method is sensitive to different conditions. In the present paper, a model-based DIF detection method for polytomous items and two observed score DIF detection methods were compared using four samples with varying numbers of examinees from large scale performance rating data. The DIF detection results revealed that there was a moderate to high similarity in the magnitudes of the three DIF statistics, $G^2$, $M^2$, and $Q^2$, within each sample and across samples except $N = 100$. The results also indicated that almost the same sets of items were identified as DIF items within each sample. One point that became clear when analyzing these samples was that statistical testing of DIF might not be a good idea when a large sample, say $N \geq 10,000$, was used. When $N = 105,731$, the three DIF detection methods identified nearly all items as statistically significant DIF items. When $N = 10,000$, more than half of the items were identified as DIF items. This sensitivity of statistical testing of DIF toward the large sample size gives a test developer a pain in the neck.

There seem to be two ways to relieve the sensitivity to the sample size in a DIF analysis. One obvious way is not to use a large sample size in a DIF analysis. Instead we may use portions of randomly sampled reference and focal groups of examinees (as we did in this study). Based on Type I error and power studies (e.g., Ankenmann et al., 1999; Zwick et al., 1993a) and parameter recovery studies for the model-based case (e.g., Reise & Yu, 1990), we may choose an appropriate sample size for the DIF analysis. This study does not offer any specific number for this, but $N = 1,000$ seems to be a good starting place. The second and more gratifying solution is to use descriptive DIF measures because we then can use all information contained in the data. Both model-based and observed score indices of standardized impact seem to be a potentially useful means of measuring and describing the amount of DIF. Note that the area between two item true score functions (or two empirical expected score functions) provides sample-independent measure of impact. When the same area is weighted by the proficiency distribution (or the relative frequency) of the focal group, the standardized impact is obtained. As Wainer (1993) indicated, this amount of DIF is sample dependent.

When we would like to perform classifications using derived scores or summed scores under the criterion-reference testing framework, the plots of not-yet-integrated model-based
index of standardized impact,

\[ \left[ T_R(\theta) - T_F(\theta) \right] g_F(\theta), \]  

(38)

and of the un-summed observed score index of standardized impact,

\[ \left[ E_R(Y|X = l) - E_F(Y|X = l) \right] \frac{N_{FI}}{N}, \]  

(39)

will be useful because these will demonstrate the amount of differential impact for the specific proficiency levels or summed scores that are used as cutoff scores. In addition, visual displays of item true score functions, proficiency distributions, and indices of standardized impact can facilitate data interpretation. The visual inspection of the item true score functions seems to be especially important as it will enhance the interpretability of \( T(3) \) and \( T'(3) \). When the amount of cancellation due to nonuniform DIF is of interest, instead of Equations 28 and 36, we may use

\[ T(3) = \int_{-\infty}^{\infty} \left| T_R(\theta) - T_F(\theta) \right| G_F(\theta) \]  

(40)

and

\[ T'(3) = \sum_{l=0}^{L} \left| E_R(Y|X = l) - E_F(Y|X = l) \right| \frac{N_{FI}}{N}. \]  

(41)

Comparisons of these with \( T(1) \) and \( T'(1) \) will provide the information with regard to the cancellation effect.

The observed score indices of standardized impact might be less suspect to the potential side effects of model misfit than the model-based indices of standardized impact. This is because there may be a confounding effect of model misfit and DIF in a model-based DIF detection method (Ankenmann et al., 1999; Dorans & Schmitt, 1991). It can be noted that if both model-based and observed score indices were obtained, then we may separate model misfit from DIF by comparing \( T(1) \) and \( T'(1) \) or \( T(3) \) and \( T'(3) \). This separation may be more obvious when the partial credit model is used in calibration because the same summed scores yield the same proficiency estimates under the partial credit model.

Due to a lack of understanding and experience using the model-based and observed score indices of standardized impact, studies are needed to investigate various applications of these indices to real data with an eye toward examining what is in fact measured by each index. It also would be useful to explore the role of these indices in the context of studies that deal with sample size and statistical power in DIF detection.

Finally, and perhaps most importantly, the justification of the tentative five percent or .05 cutoff value of the indices of the standardized impact was mainly based on the statements
of the original contributors of the SMD (Dorans & Holland, 1993; Dorans & Kulick, 1986; Dorans & Schmitt, 1991). There still remain important issues to be examined. One might state, as Rosnow and Rosenthal (1989) did in the context of the Type I error assignment in hypothesis testing, “Surely, God loves the .06 nearly as much as the .05,” which elicited “Amen” from Cohen (1990). Hence, we still need to establish firmly from the accumulation of experience a bad-enough value (which is a counterpart of the good-enough value in Serlin & Lapsley, 1993), $\Delta_s$, where $s$ indicates the smallest difference that would constitute a nontrivial DIF effect. Here, I am just uttering/paraphrasing:

I do not know whether God loves .06 as much as .05; but to myself I and many others seem to have been fond of .05 or 1/20, because we have been told hither and thither that the size of a just noticeable difference interval, called $\Delta S$, is proportional to the size of the stimulus, $S$, (i.e., Weber’s law, for example, $\Delta S/S$ is roughly .05 for lifted weights; Calfee, 1975), whilst the real meaning of this in DIF lay all undiscovered before us.
References


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.


Table 1

Data for the ith Level of the Matching Variable

<table>
<thead>
<tr>
<th>Group</th>
<th>$Y_1$</th>
<th>...</th>
<th>$Y_k$</th>
<th>...</th>
<th>$Y_K$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focal</td>
<td>$A_{1l}$</td>
<td>...</td>
<td>$A_{kl}$</td>
<td>...</td>
<td>$A_{Kl}$</td>
<td>$N_{Fl}$</td>
</tr>
<tr>
<td>Reference</td>
<td>$B_{1l}$</td>
<td>...</td>
<td>$B_{kl}$</td>
<td>...</td>
<td>$B_{Kl}$</td>
<td>$N_{Rl}$</td>
</tr>
<tr>
<td>Total</td>
<td>$M_{1l}$</td>
<td>...</td>
<td>$M_{kl}$</td>
<td>...</td>
<td>$M_{Kl}$</td>
<td>$T_i$</td>
</tr>
</tbody>
</table>
Table 2
Summary Statistics for Male (Focal), Female (Reference), and Total Group

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N = 105,731</th>
<th>N = 10,000</th>
<th>N = 1,000</th>
<th>N = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>Total</td>
<td>Male</td>
</tr>
<tr>
<td>No. of Examinees</td>
<td>55,017</td>
<td>50,714</td>
<td>105,731</td>
<td>5,000</td>
</tr>
<tr>
<td>Mean</td>
<td>11.41</td>
<td>12.43</td>
<td>11.90</td>
<td>11.52</td>
</tr>
<tr>
<td>SD</td>
<td>4.21</td>
<td>4.06</td>
<td>4.17</td>
<td>4.12</td>
</tr>
<tr>
<td>Range</td>
<td>0–20</td>
<td>0–20</td>
<td>0–20</td>
<td>0–20</td>
</tr>
<tr>
<td>Alpha</td>
<td>.83</td>
<td>.83</td>
<td>.83</td>
<td>.82</td>
</tr>
</tbody>
</table>
Table 3
Item Mean, Standard Deviation (SD), and Correlation (r) with Corrected Total Score for N = 105,731

<table>
<thead>
<tr>
<th>Item</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>r</td>
</tr>
<tr>
<td>1</td>
<td>0.96</td>
<td>.64</td>
<td>.54</td>
</tr>
<tr>
<td>2</td>
<td>1.11</td>
<td>.72</td>
<td>.54</td>
</tr>
<tr>
<td>3</td>
<td>0.74</td>
<td>.50</td>
<td>.47</td>
</tr>
<tr>
<td>4</td>
<td>0.79</td>
<td>.61</td>
<td>.64</td>
</tr>
<tr>
<td>5</td>
<td>1.16</td>
<td>.75</td>
<td>.41</td>
</tr>
<tr>
<td>6</td>
<td>1.76</td>
<td>.58</td>
<td>.49</td>
</tr>
<tr>
<td>7</td>
<td>1.22</td>
<td>.81</td>
<td>.66</td>
</tr>
<tr>
<td>8</td>
<td>1.13</td>
<td>.72</td>
<td>.57</td>
</tr>
<tr>
<td>9</td>
<td>1.58</td>
<td>.62</td>
<td>.46</td>
</tr>
<tr>
<td>10</td>
<td>0.96</td>
<td>.71</td>
<td>.43</td>
</tr>
</tbody>
</table>
Table 4

*Latent Roots of the Correlation Matrix*

<table>
<thead>
<tr>
<th>Order</th>
<th>$N = 105,731$</th>
<th>$N = 10,000$</th>
<th>$N = 1,000$</th>
<th>$N = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.05</td>
<td>3.98</td>
<td>3.92</td>
<td>4.47</td>
</tr>
<tr>
<td>2</td>
<td>0.94</td>
<td>0.95</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>0.85</td>
<td>0.87</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>4</td>
<td>0.73</td>
<td>0.74</td>
<td>0.76</td>
<td>0.74</td>
</tr>
<tr>
<td>5</td>
<td>0.71</td>
<td>0.71</td>
<td>0.72</td>
<td>0.68</td>
</tr>
<tr>
<td>6</td>
<td>0.64</td>
<td>0.65</td>
<td>0.66</td>
<td>0.59</td>
</tr>
<tr>
<td>7</td>
<td>0.63</td>
<td>0.64</td>
<td>0.61</td>
<td>0.55</td>
</tr>
<tr>
<td>8</td>
<td>0.58</td>
<td>0.58</td>
<td>0.57</td>
<td>0.44</td>
</tr>
<tr>
<td>9</td>
<td>0.51</td>
<td>0.52</td>
<td>0.48</td>
<td>0.35</td>
</tr>
<tr>
<td>10</td>
<td>0.36</td>
<td>0.36</td>
<td>0.37</td>
<td>0.31</td>
</tr>
<tr>
<td>Item</td>
<td>Compact Model</td>
<td>Augmented Model</td>
<td>Reference/Anchor Item</td>
<td>Focal Item</td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>----------------</td>
<td>-----------------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td>( a(s.e.) )</td>
<td>( b(s.e.) )</td>
<td>( a(s.e.) )</td>
<td>( b(s.e.) )</td>
</tr>
<tr>
<td>1</td>
<td>1.52(.01)</td>
<td>-1.33(.01)</td>
<td>1.27(.01)</td>
<td>1.48(.02)</td>
</tr>
<tr>
<td>2</td>
<td>1.42(.01)</td>
<td>-1.45(.01)</td>
<td>.52(.01)</td>
<td>1.42(.01)</td>
</tr>
<tr>
<td>3</td>
<td>1.36(.01)</td>
<td>-1.16(.01)</td>
<td>2.92(.02)</td>
<td>1.36(.01)</td>
</tr>
<tr>
<td>4</td>
<td>2.75(.02)</td>
<td>-1.71(.01)</td>
<td>1.39(.01)</td>
<td>2.76(.02)</td>
</tr>
<tr>
<td>5</td>
<td>1.43(.01)</td>
<td>-1.42(.02)</td>
<td>.61(.01)</td>
<td>1.43(.01)</td>
</tr>
<tr>
<td>6</td>
<td>2.61(.02)</td>
<td>-2.17(.01)</td>
<td>-1.46(.01)</td>
<td>2.61(.02)</td>
</tr>
<tr>
<td>7</td>
<td>2.48(.02)</td>
<td>-1.64(.01)</td>
<td>.08(.01)</td>
<td>2.62(.02)</td>
</tr>
<tr>
<td>8</td>
<td>1.71(.01)</td>
<td>-1.30(.01)</td>
<td>.53(.01)</td>
<td>1.71(.01)</td>
</tr>
<tr>
<td>9</td>
<td>1.18(.01)</td>
<td>-2.84(.02)</td>
<td>-.77(.01)</td>
<td>1.18(.01)</td>
</tr>
<tr>
<td>10</td>
<td>1.09(.01)</td>
<td>-1.51(.01)</td>
<td>.96(.01)</td>
<td>1.09(.01)</td>
</tr>
</tbody>
</table>

*The compact model yielded \( \mu_F(s.e.) = -14(.01) \) and \(-2\log L = 73622.1\).*
<table>
<thead>
<tr>
<th>Item</th>
<th>Reference</th>
<th>Focal</th>
<th>Augmented Model</th>
<th>(-2 \log L)</th>
<th>(G^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a_R) (s.e.)</td>
<td>(b_{1R}) (s.e.)</td>
<td>(b_{2R}) (s.e.)</td>
<td>(a_F) (s.e.)</td>
<td>(b_{1F}) (s.e.)</td>
</tr>
<tr>
<td>1</td>
<td>1.48 (.02)</td>
<td>-1.43 (.02)</td>
<td>1.29 (.01)</td>
<td>1.54 (.02)</td>
<td>-1.26 (.02)</td>
</tr>
<tr>
<td>2</td>
<td>1.43 (.02)</td>
<td>-1.49 (.02)</td>
<td>0.47 (.01)</td>
<td>1.40 (.01)</td>
<td>-1.41 (.01)</td>
</tr>
<tr>
<td>3</td>
<td>1.37 (.02)</td>
<td>-1.31 (.02)</td>
<td>2.85 (.03)</td>
<td>1.33 (.02)</td>
<td>-1.04 (.01)</td>
</tr>
<tr>
<td>4</td>
<td>2.82 (.03)</td>
<td>-0.72 (.01)</td>
<td>1.40 (.01)</td>
<td>2.70 (.03)</td>
<td>-0.70 (.01)</td>
</tr>
<tr>
<td>5</td>
<td>0.96 (.01)</td>
<td>-1.81 (.03)</td>
<td>0.65 (.02)</td>
<td>0.91 (.01)</td>
<td>-1.82 (.03)</td>
</tr>
<tr>
<td>6</td>
<td>1.80 (.03)</td>
<td>-2.16 (.02)</td>
<td>1.44 (.02)</td>
<td>1.83 (.02)</td>
<td>-2.17 (.02)</td>
</tr>
<tr>
<td>7</td>
<td>2.70 (.03)</td>
<td>-0.90 (.01)</td>
<td>0.17 (.00)</td>
<td>2.65 (.02)</td>
<td>-0.99 (.01)</td>
</tr>
<tr>
<td>8</td>
<td>1.77 (.02)</td>
<td>-1.23 (.01)</td>
<td>0.56 (.01)</td>
<td>1.69 (.01)</td>
<td>-1.35 (.01)</td>
</tr>
<tr>
<td>9</td>
<td>1.23 (.02)</td>
<td>-2.71 (.03)</td>
<td>-0.75 (.01)</td>
<td>1.11 (.01)</td>
<td>-2.95 (.03)</td>
</tr>
<tr>
<td>10</td>
<td>1.14 (.01)</td>
<td>-1.76 (.02)</td>
<td>0.70 (.01)</td>
<td>1.01 (.01)</td>
<td>-1.32 (.02)</td>
</tr>
</tbody>
</table>
Table 7
Likelihood Ratio Statistics $G^2$, Mantel Statistics $M^2$, and Generalized Mantel Haenszel Statistics $Q^2$

<table>
<thead>
<tr>
<th>Item</th>
<th>N = 105,731</th>
<th>N = 10,000</th>
<th>N = 1,000</th>
<th>N = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$G^2$</td>
<td>$M^2$</td>
<td>$Q^2$</td>
<td>$G^2$</td>
</tr>
<tr>
<td>1</td>
<td>116.9</td>
<td>2.42</td>
<td>44.50</td>
<td>8.1</td>
</tr>
<tr>
<td>2</td>
<td>77.3</td>
<td>12.88</td>
<td>14.36</td>
<td>8.0</td>
</tr>
<tr>
<td>3</td>
<td>474.6</td>
<td>314.62</td>
<td>385.62</td>
<td>10.6</td>
</tr>
<tr>
<td>4</td>
<td>19.7</td>
<td>8.45</td>
<td>29.41</td>
<td>.7</td>
</tr>
<tr>
<td>5</td>
<td>118.0</td>
<td>108.74</td>
<td>157.67</td>
<td>14.4</td>
</tr>
<tr>
<td>6</td>
<td>26.3</td>
<td>41.54</td>
<td>44.76</td>
<td>3.8</td>
</tr>
<tr>
<td>7</td>
<td>901.2</td>
<td>629.65</td>
<td>747.86</td>
<td>68.5</td>
</tr>
<tr>
<td>8</td>
<td>146.1</td>
<td>243.45</td>
<td>243.86</td>
<td>16.6</td>
</tr>
<tr>
<td>9</td>
<td>38.0</td>
<td>29.28</td>
<td>69.47</td>
<td>17.2</td>
</tr>
<tr>
<td>10</td>
<td>2101.0</td>
<td>1743.54</td>
<td>1744.32</td>
<td>205.7</td>
</tr>
</tbody>
</table>

The .01 level critical values are $\chi^2_{df=3} = 11.34$ for $G^2$, $\chi^2_{df=1} = 6.63$ for $M^2$, and $\chi^2_{df=2} = 9.21$ for $Q^2$. 
Table 8

Spearman’s Correlations Among DIF Indices

<table>
<thead>
<tr>
<th>Sample</th>
<th>DIF Index</th>
<th>( N = 105,731 )</th>
<th>( N = 10,000 )</th>
<th>( N = 1,000 )</th>
<th>( N = 100 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( G^2 )</td>
<td>( M^2 )</td>
<td>( Q^2 )</td>
<td>( G^2 )</td>
<td>( M^2 )</td>
</tr>
<tr>
<td></td>
<td>( N = 105,731 )</td>
<td>( N = 10,000 )</td>
<td>( N = 1,000 )</td>
<td>( N = 100 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( G^2 )</td>
<td>.83</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( M^2 )</td>
<td>.87</td>
<td>.94</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( Q^2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( N = 10,000 )</td>
<td>( N = 1,000 )</td>
<td>( N = 100 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( G^2 )</td>
<td>.78</td>
<td>.73</td>
<td>.83</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>( M^2 )</td>
<td>.82</td>
<td>.96</td>
<td>.93</td>
<td>.75</td>
</tr>
<tr>
<td></td>
<td>( Q^2 )</td>
<td>.83</td>
<td>.86</td>
<td>.94</td>
<td>.79</td>
</tr>
<tr>
<td></td>
<td>( N = 1,000 )</td>
<td>( N = 100 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( G^2 )</td>
<td>.29</td>
<td>.56</td>
<td>.52</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>( M^2 )</td>
<td>.24</td>
<td>.59</td>
<td>.51</td>
<td>.27</td>
</tr>
<tr>
<td></td>
<td>( Q^2 )</td>
<td>.22</td>
<td>.61</td>
<td>.53</td>
<td>.21</td>
</tr>
<tr>
<td></td>
<td>( N = 100 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( G^2 )</td>
<td>-.21</td>
<td>-.23</td>
<td>-.23</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>( M^2 )</td>
<td>-.07</td>
<td>-.14</td>
<td>-.02</td>
<td>-.16</td>
</tr>
<tr>
<td></td>
<td>( Q^2 )</td>
<td>-.03</td>
<td>-.07</td>
<td>.14</td>
<td>.22</td>
</tr>
<tr>
<td>Item</td>
<td>Model-Based Index</td>
<td>Observed-Score Index</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------------------</td>
<td>----------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T(1)</td>
<td>T(2)</td>
<td>T(3)</td>
<td>T(4)</td>
<td>T'(1)</td>
</tr>
<tr>
<td>1</td>
<td>.0188</td>
<td>1034.59</td>
<td>.0010</td>
<td>56.38</td>
<td>.0067</td>
</tr>
<tr>
<td>2</td>
<td>.0433</td>
<td>2381.63</td>
<td>.0020</td>
<td>107.90</td>
<td>.0131</td>
</tr>
<tr>
<td>3</td>
<td>.0555</td>
<td>3052.95</td>
<td>.0034</td>
<td>189.56</td>
<td>.0475</td>
</tr>
<tr>
<td>4</td>
<td>-.0017</td>
<td>-93.08</td>
<td>.0001</td>
<td>7.54</td>
<td>-.0070</td>
</tr>
<tr>
<td>5</td>
<td>-.0072</td>
<td>-398.08</td>
<td>.0001</td>
<td>5.66</td>
<td>-.0394</td>
</tr>
<tr>
<td>6</td>
<td>.0106</td>
<td>584.12</td>
<td>.0002</td>
<td>8.92</td>
<td>-.0179</td>
</tr>
<tr>
<td>7</td>
<td>-.0547</td>
<td>-3008.52</td>
<td>.0054</td>
<td>294.62</td>
<td>-.0792</td>
</tr>
<tr>
<td>8</td>
<td>-.0160</td>
<td>-879.32</td>
<td>.0005</td>
<td>29.91</td>
<td>-.0521</td>
</tr>
<tr>
<td>9</td>
<td>.0134</td>
<td>737.16</td>
<td>.0002</td>
<td>13.72</td>
<td>-.0196</td>
</tr>
<tr>
<td>10</td>
<td>.1785</td>
<td>9819.36</td>
<td>.0333</td>
<td>1832.03</td>
<td>.1479</td>
</tr>
</tbody>
</table>
Table 10
Cross Classification of Item 10 Score by Summed Score for Male and Female

<table>
<thead>
<tr>
<th>Male-Focal Group</th>
<th>Summed Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Score</td>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20</td>
</tr>
<tr>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20</td>
<td></td>
</tr>
<tr>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>475 527 725 1025 1319 1743 2088 2490 2870 3257 3794 4484 5115 5574 5905 6083 6242 6462 6666 6864</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Female-Reference Group</th>
<th>Summed Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Score</td>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20</td>
</tr>
<tr>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20</td>
<td></td>
</tr>
<tr>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>240 245 315 391 480 584 576 658 624 635 641 636 589 500 345 197 64 27 9 0 0</td>
</tr>
</tbody>
</table>

BEST COPY AVAILABLE

36 38
Figure Captions

Figure 1. Latent roots in order of size for $N = 105,731$, $N = 10,000$, $N = 1,000$, and $N = 100$.

Figure 2. Illustration of the calculation of the model-based index of standardized impact.

Figure 3. Illustration of the calculation of the observed score index of standardized impact.
Appendix

Zwick and Thayer (1996) presented two variances for the SMD. Using the notation in Table 1, the two variances are presented below. It should be noted that the first is based on the hypergeometric model and known to provide better standard errors (Zwick & Thayer, 1996).

The first variance of the SMD were given as

\[
\text{Var}_H(\text{SMD}) = \sum_{l=0}^{L} w_{Fl} \left( \frac{1}{N_{Fl}} + \frac{1}{N_{Rl}} \right)^2 \text{Var}_H(F_l),
\]

where the subscript \(H\) designates the hypergeometric framework,

\[
w_{Fl} = \frac{N_{Fl}}{N_F},
\]

\[
F_l = \sum_{k=1}^{K} A_{kl} Y_k,
\]

and \(\text{Var}_H(F_l)\) is defined in Equation 16.

The second based on the multinomial model is

\[
\text{Var}_M(\text{SMD}) = \sum_{l=0}^{L} w_{Fl}^2 \left[ \left( \frac{1}{N_{Fl}} \right)^2 \text{Var}_M(F_l) + \left( \frac{1}{N_{Rl}} \right)^2 \text{Var}_M(R_l) \right],
\]

where

\[
\text{Var}_M(F_l) = N_{Fl} \left[ \sum_{k=1}^{K} \hat{\pi}_{Flk} Y_k^2 - \left( \sum_{k=1}^{K} \hat{\pi}_{Flk} Y_k \right)^2 \right],
\]

\[
\text{Var}_M(R_l) = N_{Rl} \left[ \sum_{k=1}^{K} \hat{\pi}_{Rlk} Y_k^2 - \left( \sum_{k=1}^{K} \hat{\pi}_{Rlk} Y_k \right)^2 \right],
\]

\(\hat{\pi}_{Flk} = A_{kl}/N_{Fl}\), and \(\hat{\pi}_{Rlk} = B_{kl}/N_{Rl}\). The subscript \(M\), of course, indicates that these expressions are obtained using the multinomial model.
Author Note

The author wishes to express his gratitude to Steve Cramer for making the data available for this study, Paula J. Schwanenflugel for a helpful conversation about sex differences in the preschool period, and to Lisa A. Garberson for her helpful comments on the earlier draft.

Correspondence concerning this manuscript should be addressed to Seock-Ho Kim, Department of Educational Psychology, The University of Georgia, 325 Aderhold Hall, Athens, GA 30602-7143. Electronic mail may be sent via internet to: skim@coe.uga.edu
Title: An Investigation of the Likelihood Ratio Test, the Mantel Test, and the Generalized Mantel-Haenszel Test of OR

Author(s): Seock-Ho Kim

Corporate Source: The University of Georgia

Publication Date: 2000

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.

Level 1: Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

Level 2A: Check here for Level 2A release, permitting reproduction and dissemination in microfiche, and in electronic media for ERIC archival collection subscribers only.

Level 2B: 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 reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Seock-Ho Kim

Printed Name/Position/Title: Assistant Professor

Organization/Address: 315 Ayden Hall

Athens, GA 30602
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:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20772
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
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)