ED 441 007 TM 030 805

AUTHOR	Kim, Seock-Ho
TITLE	An Investigation of the Likelihood Ratio Test, the Mantel Test, and the Generalized Mantel-Haenszel Test of DIF.
PUB DATE	2000-04-27
NOTE	44p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Item Bias; Item Response Theory; Kindergarten; Performance Based Assessment; Primary Education; Sample Size; *Test Items
IDENTIFIERS	Graded Response Model; *Likelihood Ratio Tests; *Mantel Haenszel Procedure

ABSTRACT

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=1,00; 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

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana

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 handcapping 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 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 $k = 1, \ldots, K_j$. In Equation 1, $P^*_{jk}(\theta)$ is the boundary 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

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

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

$$TS(\theta) = \sum_{j=1}^{J} \sum_{k=1}^{K_j} y_{jk} P_{jk}(\theta), \qquad (4)$$

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

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

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

$$u_{jk} = \begin{cases} 1 & \text{if } y_j = k \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

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

$$\text{Prob}\left\{y_j = k|\theta, \boldsymbol{\xi}_j\right\} = P_{jk}(\theta) = \prod_{k=1}^{K} P_{jk}(\theta)^{u_{jk}}, \tag{7}$$

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

$$P(\mathbf{y}_l|\theta) = \prod_{j=1}^{J} \prod_{k=1}^{K} P_{jk}(\theta)^{u_{jk}}, \tag{8}$$

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

$$P(\mathbf{y}_l) = \int P(\mathbf{y}_l|\theta) g(\theta|\boldsymbol{\tau}) d\theta = \int P(\mathbf{y}_l|\theta) dG(\theta|\boldsymbol{\tau}), \tag{9}$$

where $g(\theta|\boldsymbol{\tau})$ 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. 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 response pattern $\mathbf{y}_l$ is

$$P_g(\mathbf{y}_l) = \int P(\mathbf{y}_l|\theta) g(\theta|\boldsymbol{\tau}_g) d\theta = \int P(\mathbf{y}_l|\theta) dG(\theta|\boldsymbol{\tau}_g). \tag{10}$$

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

$$\prod_{g=1}^{G} \prod_{l=1}^{K^J} P_g(\mathbf{y}_l)^{r_g(\mathbf{y}_l)}, \tag{11}$$

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 θ, 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 $\sigma_R^2 = \sigma_F^2$, while μ_F for the focal group is estimated from data. Then,

$$-2 \log L = -2 \sum_{g=1}^{G} \sum_{l=1}^{K^J} r_g(\mathbf{y}_l) \log \left[\frac{N_g \hat{P}_g(\mathbf{y}_l)}{r_g(\mathbf{y}_l)} \right], \tag{12}$$

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.]

The likelihood ratio test statistic can be written as

$$G^2 = -2 \log L_{\mathrm{C}} - (-2 \log L_{\mathrm{A}}) \tag{13}$$

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 three categories is tested, G^2 is distributed as a χ^2 with 3 degrees of freedom.

Mantal 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 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 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 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 for the focal group,

$$\sum_{k=1}^{K} A_{kl} Y_k, \tag{14}$$

has its expectation and variance defined as

$$E\left(\sum_{k=1}^{K} A_{kl} Y_k\right) = \frac{N_{Fl} \sum_{k=1}^{K} M_{kl} Y_k}{T_l} \tag{15}$$

and

$$\text{Var}\left(\sum_{k=1}^{K} A_{kl} Y_k\right) = \frac{N_{Fl} N_{Rl}}{T_l^2 (T_l - 1)} \left[T_l \sum_{k=1}^{K} M_{kl} Y_k^2 - \left(\sum_{k=1}^{K} M_{kl} Y_k\right)^2\right]. \tag{16}$$

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

$$\frac{\left[\sum_{k=1}^{K} A_{kl} Y_k - E\left(\sum_{k=1}^{K} A_{kl} Y_k\right)\right]^2}{\text{Var}\left(\sum_{k=1}^{K} A_{kl} Y_k\right)} \tag{17}$$

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 lth 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

$$M^2 = \frac{\left[\sum_{l=1}^{L}\sum_{k=1}^{K} A_{kl}Y_k - \sum_{l=1}^{L} E\left(\sum_{k=1}^{K} A_{kl}Y_k\right)\right]^2}{\sum_{l=1}^{L} \mathrm{Var}\left(\sum_{k=1}^{K} A_{kl}Y_k\right)}. \tag{18}$$

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

$$\mathbf{a}_l = (A_{1l}, \ldots, A_{kl}, \ldots, A_{(K-1)l})' \tag{19}$$

which has expectation

$$E(\mathbf{a}_l) = N_{Fl}\mathbf{m}_l/T_l \tag{20}$$

and variance

$$\mathbf{V}_l = \frac{N_{Fl}N_{Rl}}{T_l^2(T_l - 1)}\left[T_l \mathrm{diag}(\mathbf{m}_l) - \mathbf{m}_l\mathbf{m}_l'\right], \tag{21}$$

where

$$\mathbf{m}_l = (M_{1l}, \ldots, M_{kl}, \ldots, M_{(K-1)l})'. \tag{22}$$

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

$$\left[\mathbf{a}_l - E(\mathbf{a}_l)\right]' \mathbf{V}_l^{-1}\left[\mathbf{a}_l - E(\mathbf{a}_l)\right] \tag{23}$$

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

$$Q^2 = \left[\sum_{l=1}^{L}\mathbf{a}_l - \sum_{l=1}^{L}E(\mathbf{a}_l)\right]'\left[\sum_{l=1}^{L}\mathbf{V}_l\right]^{-1}\left[\sum_{l=1}^{L}\mathbf{a}_l - \sum_{l=1}^{L}E(\mathbf{a}_l)\right]. \tag{24}$$

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

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

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 $N = 100$.

Insert Table 2 about here

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.

Insert Table 3 about here

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.

Insert Table 4 and Figure 1 about here

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 $\alpha = .01$.

Insert Table 5 about here

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.

Insert Table 6 about here

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 $N = 100$.

Insert Table 7 about here

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 $\alpha = .01$.

Insert Table 8 about here

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

$$T(1) = \int_{-\infty}^{\infty} [T_R(\theta) - T_F(\theta)]\, dG_F(\theta), \tag{26}$$

$$T(2) = N_F T(1), \tag{27}$$

$$T(3) = \int_{-\infty}^{\infty} [T_R(\theta) - T_F(\theta)]^2\, dG_F(\theta), \tag{28}$$

and

$$T(4) = N_F T(3), \tag{29}$$

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, 1088; 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 $-.14$ and variance 1, that is, $g_F(\theta) = N(-.14, 1)$. 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 θ 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 θ, 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

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

$$T(1)/R, \tag{30}$$

where R is the range of item scores. The possible values of the index will be limited within ±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

$$T^{'}(1) = \sum_{l=0}^{L} [E_R(Y|X=l) - E_F(Y|X=l)] \frac{N_{Fl}}{N_F}, \tag{31}$$

where

$$E_R(Y_j|X=l) = \frac{\sum B_{kl}Y_k}{N_{Rl}} \tag{32}$$

and

$$E_F(Y_j|X=l) = \frac{\sum A_{kl}Y_k}{N_{Fl}} \tag{33}$$

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

$$\frac{N_{Fl}}{\sum_{l=0}^{L} N_{Fl}} = \frac{N_{FL}}{N_F} \tag{34}$$

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

$$T'(2) = N_F T'(1) = \sum_{l=0}^{L} \left[E_R(Y|X = l) - E_F(Y|X = l)\right] N_{Fl}, \tag{35}$$

$$T'(3) = \sum_{l=0}^{L} \left[E_R(Y|X = l) - E_F(Y|X = l)\right]^2 \frac{N_{Fl}}{N_F}, \tag{36}$$

and

$$T'(4) = N_F T'(3) = \sum_{l=0}^{L} \left[E_R(Y|X = l) - E_F(Y|X = l)\right]^2 N_{Fl}. \tag{37}$$

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 θ, 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 $\sum_l$ 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.

Insert Figure 3 and Table 10 about here

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 ± 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 θ 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,

$$[T_R(\theta) - T_F(\theta)]\, g_F(\theta), \tag{38}$$

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

$$[E_R(Y|X = l) - E_F(Y|X = l)]\,\frac{N_{Fl}}{N}, \tag{39}$$

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

$$T(3) = \int_{-\infty}^{\infty} |T_R(\theta) - T_F(\theta)|\, G_F(\theta) \tag{40}$$

and

$$T^{'}(3) = \sum_{l=0}^{L} |E_R(Y|X = l) - E_F(Y|X = l)|\,\frac{N_{Fl}}{N}. \tag{41}$$

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), Δ_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 ΔS, 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

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

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

American Psychological Association (APA). (1988). *Code of fair testing practices in education.* Washington, DC: Author.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36,* 279–300.

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

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

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.

Calfee, R. C. (1975). *Human experimental psychology.* New York: Holt, Rinehart and Winston.

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.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (Research Rep. No. RR-91-47). Princeton, NJ: Educational Testing Service.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessment. *Applied Psychological Measurement, 4,* 289–303.

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

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

Kim, S.-H. & Cohen, A. S. (1988). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22,* 345–355.

Linn, R. L., Levine, M. V., Hasting, C. N., & Wardrop, J. L. (1980). *An investigation of item bias in a test of reading comprehension* (Tech. Rep. No. 163). Center for the Study of Reading, University of Illinois at Urbana-Champaign.

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.

Miller, T. A., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107–122.

Muthén, B. O. (1988). LISCOMP: Analysis of linear structural equations with a comprehensive measurement model (2nd ed.) [Computer program]. Mooresville, IN: Scientific Software.

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.

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.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23–37.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495–502.

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.

Ramsay, J. O. (1982). When data are functions. *Psychometrika, 47,* 379–396.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133–144.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

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

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometric Monographs,* No. 17.

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

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer-Verlag.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.

Shaffer, D. R. (2000). *Social and personality development* (4th ed.). Belmont, CA: Wadsworth.

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

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory (Version 6.0) [Computer program]. Chicago: Scientific Software.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.

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

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

Wainer, H., & Braun, H. I. (1988). *Test validity.* Hillsdale, NJ: Erlbaum.

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.

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

Zwick, R., Donoghue, J. R., & Grima, A. (1993b). *Assessment of differential item functioning for performance tests* (Research Rep. No. RR-93-14). Princeton, NJ: Educational Testing Service.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21,* 187–201.

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 Male (Focal), Female (Reference), and Total Group

	$N = 105,731$			$N = 10,000$			$N = 1,000$			$N = 100$		
Statistic	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total
No. of Examinees	55,017	50,714	105,731	5,000	5,000	10,000	500	500	1,000	50	50	100
Mean	11.41	12.43	11.90	11.52	12.40	12.48	11.60	12.12	11.86	11.76	11.72	11.74
SD	4.21	4.06	4.17	4.12	4.06	4.12	4.07	4.07	4.07	4.47	4.30	4.36
Range	0–20	0–20	0–20	0–20	0–20	0–20	0–20	0–20	0–20	1–19	1–20	1–20
Alpha	.83	.83	.83	.82	.83	.83	.82	.83	.82	.86	.86	.86

Table 3

Item Mean, Standard Deviation (SD), and Correlation (r) with Corrected Total Score for N = 105,731

	Male			Female			Total		
Item	Mean	SD	*r*	Mean	SD	*r*	Mean	SD	*r*
1	0.96	.64	.54	1.06	.61	.52	1.21	.63	.53
2	1.11	.72	.54	1.25	.71	.54	1.18	.72	.54
3	0.74	.50	.47	0.85	.47	.45	0.80	.49	.47
4	0.79	.61	.64	0.89	.60	.65	0.84	.61	.64
5	1.16	.75	.41	1.22	.72	.42	1.19	.73	.42
6	1.76	.58	.49	1.81	.52	.46	1.78	.55	.48
7	1.22	.81	.66	1.29	.78	.66	1.25	.80	.66
8	1.13	.72	.57	1.20	.71	.57	1.16	.72	.57
9	1.58	.62	.46	1.64	.59	.47	1.61	.60	.47
10	0.96	.71	.43	1.22	.69	.46	1.08	.71	.45

Table 4

Latent Roots of the Correlation Matrix

	Sample Size			
Order	$N = 105,731$	$N = 10,000$	$N = 1,000$	$N = 100$
1	4.05	3.98	3.92	4.47
2	.94	.95	.97	.98
3	.85	.87	.94	.89
4	.73	.74	.76	.74
5	.71	.71	.72	.68
6	.64	.65	.66	.59
7	.63	.64	.61	.55
8	.58	.58	.57	.44
9	.51	.52	.48	.35
10	.36	.36	.37	.31

Table 5

Item Parameter Estimates and Standard Errors (s.e.) from the Compact and Augmented Models and the Likelihood Ratio Statistic G^2 for Item 1

				Augmented Model									
	Compact Model[a]			Reference/Anchor Item			Focal						
Item	a(s.e.)	b_1(s.e.)	b_2(s.e.)	a_R(s.e.)	b_{1R}(s.e.)	b_{2R}(s.e.)	a_F(s.e.)	b_{1F}(s.e.)	b_{2F}(s.e.)	$\hat{\mu}_F$(s.e.)	$-2 \log L$	G^2
1	1.51(.01)	−1.33(.01)	1.27(.01)	1.48(.02)	−1.43(.02)	1.29(.01)	1.54(.02)	−1.26(.02)	1.25(.01)	−.14(.01)	73505.2	116.9
2	1.42(.01)	−1.45(.01)	.52(.01)	1.42(.01)	−1.44(.01)	.53(.01)						
3	1.36(.01)	−1.16(.01)	2.92(.02)	1.36(.01)	−1.16(.01)	2.92(.02)						
4	2.75(.02)	−.71(.01)	1.39(.01)	2.76(.02)	−.71(.01)	1.39(.01)						
5	.93(.01)	−1.82(.02)	.61(.01)	.93(.01)	−1.82(.02)	.61(.01)						
6	1.81(.02)	−2.17(.01)	−1.46(.01)	1.81(.02)	−2.17(.01)	−1.46(.01)						
7	2.61(.02)	−.94(.01)	.08(.01)	2.62(.02)	−.94(.01)	.05(.00)						
8	1.71(.01)	−1.30(.01)	.53(.01)	1.71(.01)	−1.30(.01)	.54(.01)						
9	1.16(.01)	−2.84(.02)	−.77(.01)	1.16(.01)	−2.84(.02)	−.77(.01)						
10	1.09(.01)	−1.51(.01)	.96(.01)	1.09(.01)	−1.51(.01)	.96(.01)						

[a] The compact model yielded $\hat{\mu}_F$(s.e.) $= -.14(.01)$ and $-2 \log L = 73622.1$.

Table 6

Item Parameter Estimates and Standard Errors (s.e.) from the Augmented Models, $-2\log L$*, and* G^2

	Augmented Model								
	Reference			Focal					
Item	a_R(s.e.)	b_{1R}(s.e.)	b_{2R}(s.e.)	a_F(s.e.)	b_{1F}(s.e.)	b_{2F}(s.e.)	$\hat{\mu}_F$(s.e.)	$-2\log L$	G^2
1	1.48(.02)	−1.43(.02)	1.29(.01)	1.54(.02)	−1.26(.02)	1.25(.01)	−.14(.01)	73505.2	116.9
2	1.43(.02)	−1.49(.02)	.47(.01)	1.40(.01)	−1.41(.01)	.59(.01)	−.13(.01)	73544.8	77.3
3	1.37(.02)	−1.31(.02)	2.85(.03)	1.33(.02)	−1.04(.01)	3.06(.04)	−.13(.01)	73147.5	474.6
4	2.82(.03)	−0.72(.01)	1.40(.01)	2.70(.03)	−.70(.01)	1.38(.01)	−.14(.01)	73602.4	19.7
5	0.96(.01)	−1.81(.03)	.65(.02)	.91(.01)	−1.82(.03)	.57(.02)	−.14(.01)	73504.1	118.0
6	1.80(.03)	−2.16(.02)	1.44(.02)	1.83(.02)	−2.17(.02)	1.48(.01)	−.14(.01)	73595.8	26.3
7	2.70(.03)	−.90(.01)	.17(.00)	2.65(.02)	−.99(.01)	−.03(.01)	−.17(.01)	72720.9	901.2
8	1.77(.02)	−1.23(.01)	.56(.01)	1.69(.01)	−1.35(.01)	.50(.01)	−.15(.01)	73476.0	146.1
9	1.23(.02)	−2.71(.03)	−.75(.01)	1.11(.01)	−2.95(.03)	−.78(.01)	−.14(.01)	73584.1	38.0
10	1.14(.01)	−1.76(.02)	.70(.01)	1.01(.01)	−1.32(.02)	1.32(.02)	−.11(.01)	71521.1	2101.0

Table 7

Likelihood Ratio Statistics G^2, Mantel Statistics M^2, and Generalized Mantel Haenszel Statistics Q^2

	$N = 105,731$			$N = 10,000$			$N = 1,000$			$N = 100$		
Item	G^2	M^2	Q^2	G^2	M^2	Q^2	G^2	M^2	Q^2	G^2	M^2	Q^2
1	116.9	2.42	44.50	8.1	.00	2.61	1.1	.17	.35	1.7	1.44	2.32
2	77.3	12.88	14.36	8.0	1.28	1.29	.3	.49	.51	3.7	.06	.37
3	474.6	314.62	385.62	10.6	44.55	47.02	2.2	2.70	2.75	1.1	.11	.89
4	19.7	8.45	29.41	.7	.01	3.13	5.5	2.77	5.60	3.1	1.30	1.42
5	118.0	108.74	157.67	14.4	26.18	26.92	4.4	1.20	2.60	1.5	.34	.99
6	26.3	41.54	44.76	3.8	2.20	2.47	10.1	3.70	9.20	1.5	1.06	2.14
7	901.2	629.65	747.86	68.5	38.42	52.07	8.8	5.34	6.29	.5	.03	.41
8	146.1	243.45	243.86	16.6	22.17	22.26	1.7	.30	.60	3.6	1.01	2.06
9	38.0	29.28	69.47	17.2	11.49	18.22	1.0	1.21	1.21	6.0	.10	2.20
10	2101.0	1743.54	1744.32	205.7	181.25	181.25	20.4	16.97	17.91	5.7	2.96	3.22

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

	DIF	$N = 105,731$			$N = 10,000$			$N = 1,000$			$N = 100$		
Sample	Index	G^2	M^2	Q^2	G^2	M^2	Q^2	G^2	M^2	Q^2	G^2	M^2	Q^2
N =105,731	G^2	1.00											
	M^2	.83	1.00										
	Q^2	.87	.94	1.00									
$N = 10,000$	G^2	.78	.73	.83	1.00								
	M^2	.82	.96	.93	.75	1.00							
	Q^2	.83	.86	.94	.79	.89	1.00						
$N = 1,000$	G^2	.29	.56	.52	.18	.47	.50	1.00					
	M^2	.24	.59	.51	.27	.54	.50	.82	1.00				
	Q^2	.22	.61	.53	.21	.55	.50	.90	.96	1.00			
$N = 100$	G^2	−.21	−.23	−.23	.16	−.18	−.21	−.35	−.18	−.24	1.00		
	M^2	−.07	−.14	−.02	−.16	−.16	−.01	.44	.06	.20	.26	1.00	
	Q^2	−.03	−.07	.14	.22	-.07	.07	.26	.06	.13	.51	.77	1.00

Table 9

Four Model-Based Indices of Standardized Impact and Four Indices of Observed Score Impact

	Model-Based Index				Observed-Score Index			
Item	$T(1)$	$T(2)$	$T(3)$	$T(4)$	$T'(1)$	$T'(2)$	$T'(3)$	$T'(4)$
1	.0188	1034.59	.0010	56.38	.0067	368.78	.0005	25.70
2	.0433	2381.63	.0020	107.90	.0131	719.84	.0003	18.24
3	.0555	3052.95	.0034	189.56	.0475	2611.77	.0034	187.26
4	-.0017	-93.08	.0001	7.54	-.0070	-385.19	.0002	11.50
5	-.0072	-398.08	.0001	5.66	-.0394	-2168.00	.0019	105.31
6	.0106	584.12	.0002	8.92	-.0179	-982.90	.0006	35.43
7	-.0547	-3008.52	.0054	294.62	-.0792	-4358.73	.0086	474.94
8	-.0160	-879.32	.0005	29.91	-.0521	-2868.14	.0032	174.85
9	.0134	737.16	.0002	13.72	-.0196	-1076.74	.0019	105.24
10	.1785	9819.36	.0333	1832.03	.1479	8139.31	.0233	1281.11

Table 10

Cross Classification of Item 10 Score by Summed Score for Male and Female

Male–Focal Group

Item	Summed Score																				
Score	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	475	404	535	736	873	1064	1121	1226	1232	1177	1251	1280	1220	893	665	383	147	49	15	0	0
1	0	123	180	274	417	634	884	1092	1456	1810	2172	2648	3170	3571	3435	2818	1698	784	291	81	0
2	0	0	13	15	29	45	83	137	182	270	371	556	725	1057	1474	1904	2038	1809	1176	585	264
Total	475	527	728	1025	1319	1743	2088	2455	2870	3257	3794	4484	5115	5521	5574	5105	3883	2642	1482	666	264

Female–Reference Group

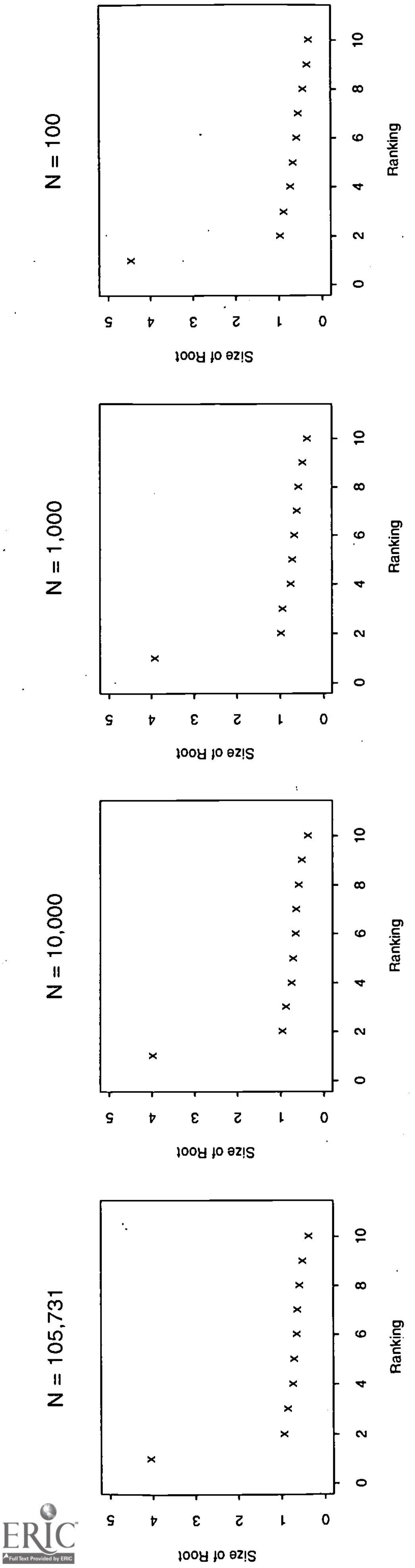
Item	Summed Score																				
Score	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	240	245	315	391	480	584	576	658	624	635	641	636	589	500	345	197	64	27	9	0	0
1	0	75	126	229	327	479	658	866	1192	1502	1894	2319	2927	3354	3095	2484	1692	716	283	64	0
2	0	0	12	16	48	56	93	131	198	332	431	650	1022	1511	2147	2764	3104	2711	1893	1057	500
Total	240	320	453	636	855	1119	1327	1655	2014	2469	2966	3605	4538	5365	5587	5445	4860	3454	2185	1121	500

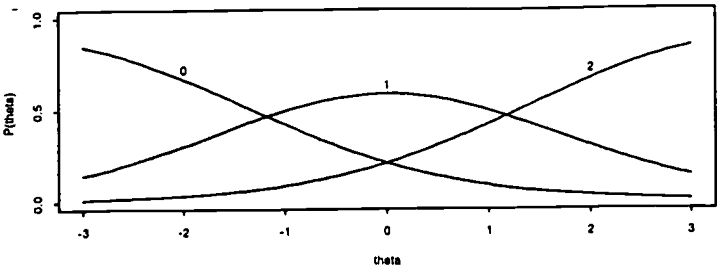
Figure Captions

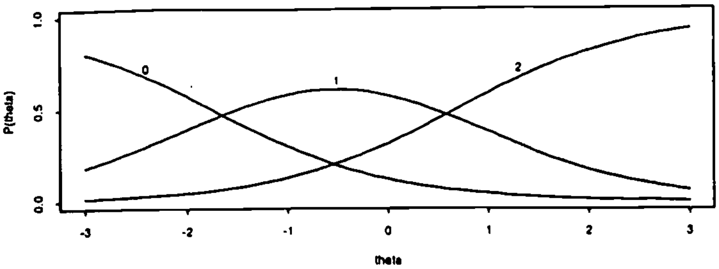
Figure 1. Latent roots in order of size for $N = 105,731$, $N = 10,000$, $N = 1,000$, and $N = 100$.

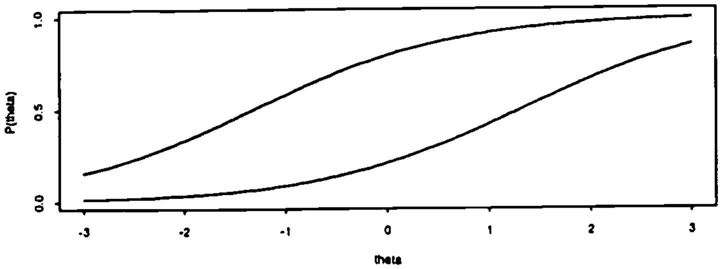
Figure 2. Illustration of the calculation of the model-based index of standardized impact.

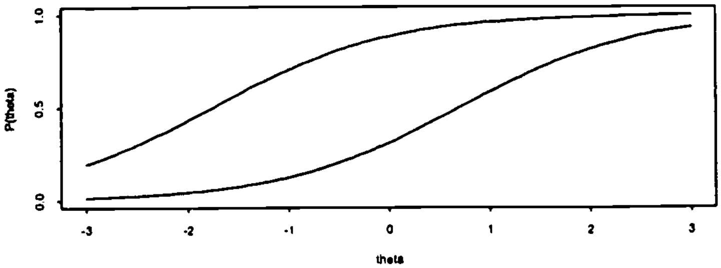
Figure 3. Illustration of the calculation of the observed score index of standardized impact.

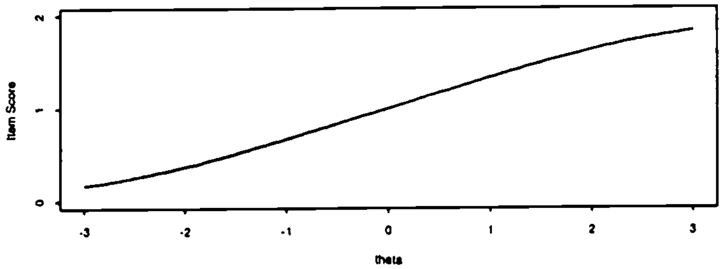

N = 100
N = 1,000
N = 10,000
N = 105,731
Size of Root
Ranking
0
1
2
3
4
5
0
2
4
6
8
10

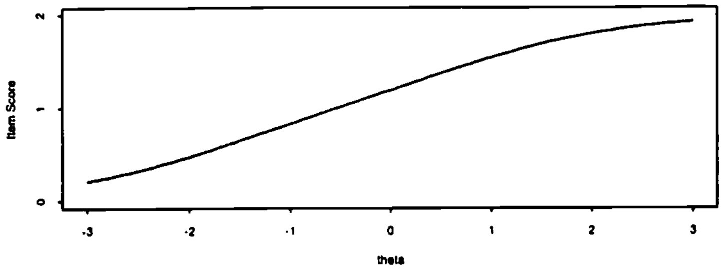

Category Response Functions for Male

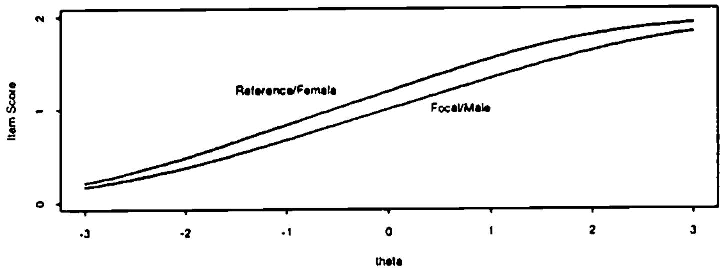

Category Response Functions for Female

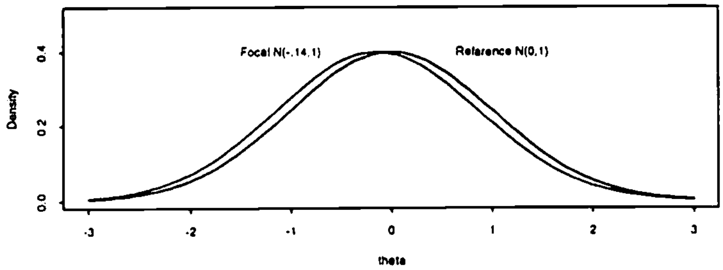

Boundary Response Functions for Male

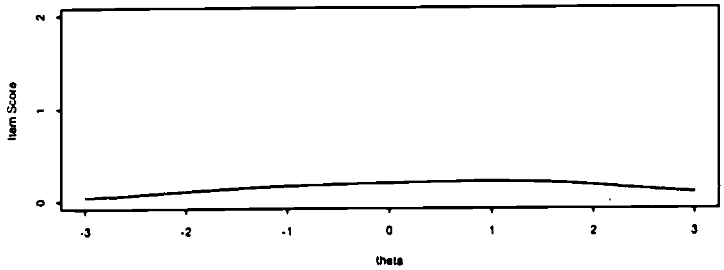

Boundary Response Functions for Female

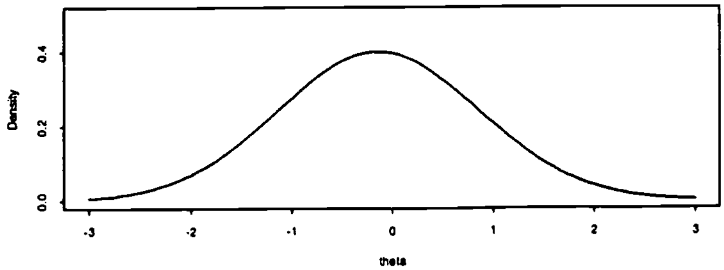

Item True Score Function for Male

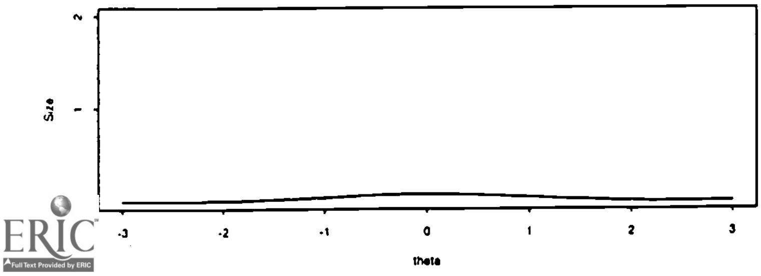

Item True Score Function for Female


Item True Score Functions

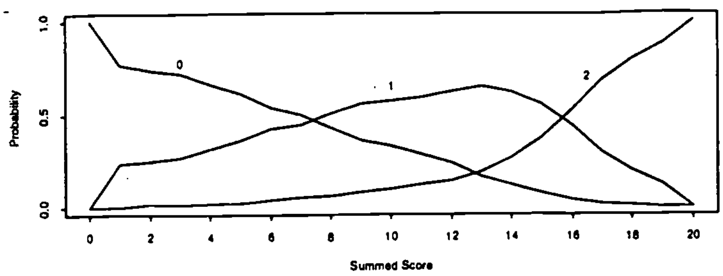

Ability Distributions


Differences in Item True Score Functions

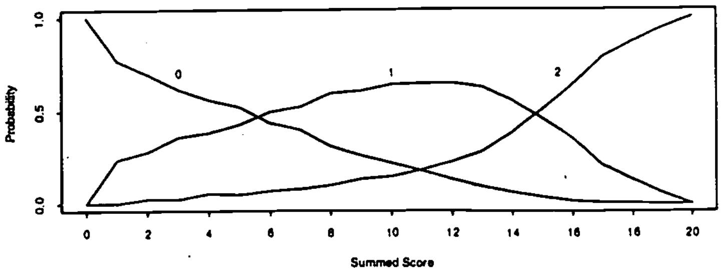

Focal Group Ability Distribution


Impact

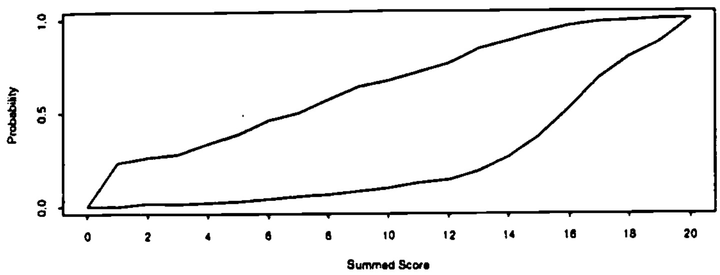
Empirical Trace Lines for Males



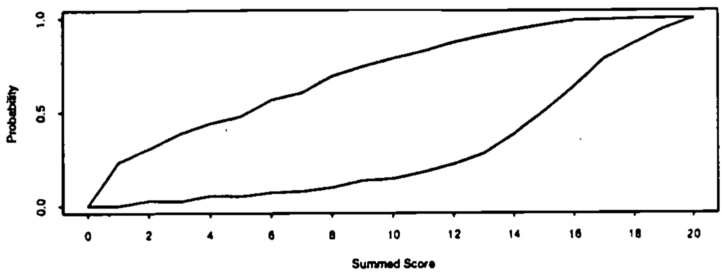
Empirical Trace Lines for Females



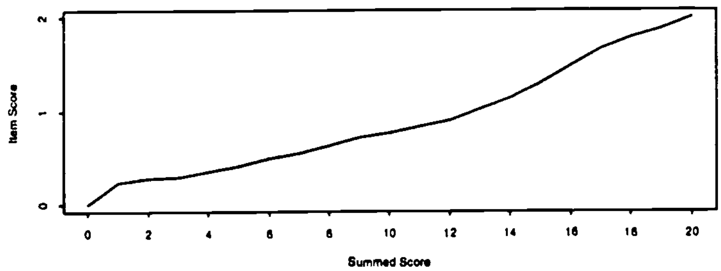
Empirical Boundary Lines for Males



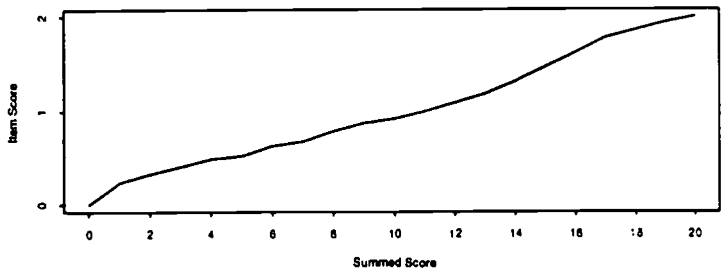
Empirical Boundary Lines for Females



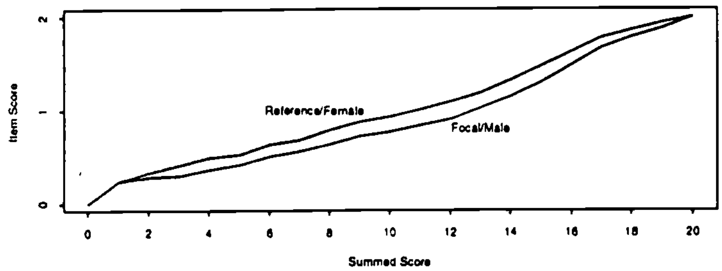
Expected Item Score for Males



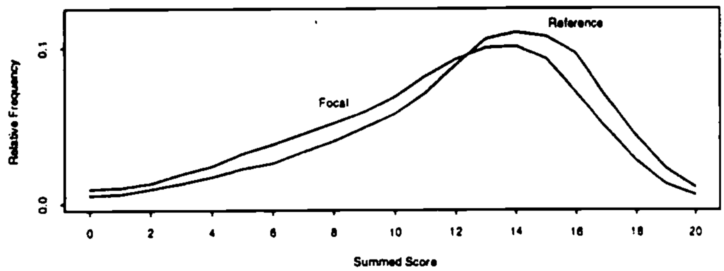
Expected Item Score for Females



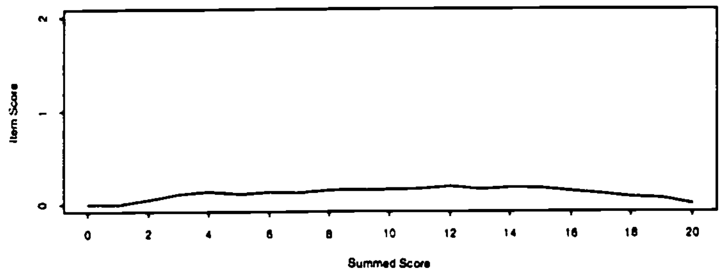
Expected Item Scores



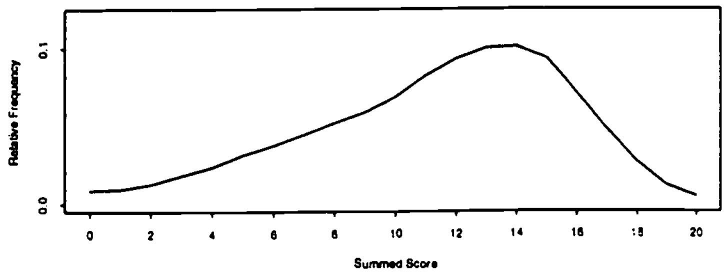
Relative Frequency of Summed Scores



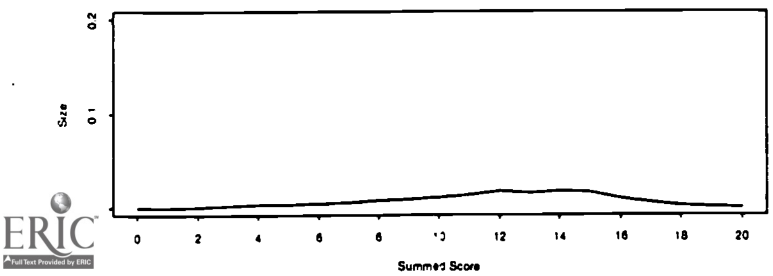
Difference in Expected Item Scores



Focal Group Relative Frequency



Observed Score 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

$$\mathrm{Var}_H(\mathrm{SMD}) = \sum_{l=0}^{L} w_{Fl} \left(\frac{1}{N_{Fl}} + \frac{1}{N_{Rl}} \right)^2 \mathrm{Var}_H(F_l), \tag{42}$$

where the subscript H designates the hypergeometric framework,

$$w_{Fl} = \frac{N_{Fl}}{N_F}, \tag{43}$$

$$F_l = \sum_{k=1}^{K} A_{kl} Y_k, \tag{44}$$

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

The second based on the multinomial model is

$$\mathrm{Var}_M(\mathrm{SMD}) = \sum_{l=0}^{L} w_{Fl}^2 \left[\left(\frac{1}{N_{Fl}} \right)^2 \mathrm{Var}_M(F_l) + \left(\frac{1}{N_{Rl}} \right)^2 \mathrm{Var}_M(R_l) \right], \tag{45}$$

where

$$\mathrm{Var}_M(F_l) = N_{Fl} \left[\sum_{k=1}^{K} \hat{\pi}_{Fkl} Y_k^2 - \left(\sum_{k=1}^{K} \hat{\pi}_{Fkl} Y_k \right)^2 \right], \tag{46}$$

$$\mathrm{Var}_M(R_l) = N_{Rl} \left[\sum_{k=1}^{K} \hat{\pi}_{Rkl} Y_k^2 - \left(\sum_{k=1}^{K} \hat{\pi}_{Rkl} Y_k \right)^2 \right], \tag{47}$$

$\hat{\pi}_{Fkl} = A_{kl}/N_{Fl}$, and $\hat{\pi}_{Rkl} = B_{kl}/N_{Rl}$. 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

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