DOCUMENT RESUME

ED 384 626 TM 023 757

AUTHOR Pommerich, Mary

TITLE Demonstrating the Utility of a Multilevel Model in the Assessment of Differential Item Functioning.

PUB DATE Apr 95

NOTE 39p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Criteria; Evaluation Methods; Grade 3; Identification; *Item Bias; Least Squares Statistics; *Measurement Techniques; Models; Prediction; Primary Education; Regression (Statistics); Research Methodology; *Test Items; True Scores; *Weighted Scores

IDENTIFIERS *Mantel Haenszel Procedure; *Multilevel Analysis; North Carolina End of Course Testing Program

ABSTRACT

When tests contain few items, observed score may not be an accurate reflection of true score, and the Mantel Haenszel (MH) statistic may perform poorly in detecting differential item functioning. Applications of the MH procedure in such situations require an alternate strategy; one such strategy is to include background variables in the matching criterion. Techniques for incorporating external information are presented here that match on a weighted score that combines the observed score and background data, using either ordinary least squares regression or a multilevel model. The regression and multilevel models were constructed using data obtained with the Grade 3 North Carolina End of Grade Mathematics Test. A simulation study was performed in which the prediction models were used to generate data, and three MH statistics were computed matching on observed scores, regression weighted scores, and multilevel weighted scores. The results showed similar performance for the regression and multilevel weighted score methods. The observed score and weighted score methods demonstrated advantages over the observed score method for test lengths of 5 and 10 items, but the improvement was small and inconsistent. Techniques for improving the performance of the weighted score methods are discussed. (Contains 21 references and 12 tables.) (SLD)

Demonstrating the Utility of a Multilevel Model in the Assessment of Differential Item Functioning[1]

Mary Pommerich
American College Testing

Abstract

Total test score is routinely employed as the matching criterion for the Mantel-Haenszel (MH) procedure for detecting differential item functioning (DIF), under the assumption that observed score is representative of true score. When tests contain few items, observed score may not be an accurate reflection of true score, and the MH statistic may perform poorly. Applications of the MH procedure in such situations require an alternate strategy; one such strategy is to include background variables in the matching criterion. Techniques for incorporating external information are presented here that match on a weighted score that combines the observed score and background data, using either ordinary least squares regression or a multilevel model. The regression and multilevel models were constructed using data obtained with the Grade 3 North Carolina End of Grade Mathematics Test. A simulation study was performed in which the prediction models were used to generate data, and three MH statistics were computed matching on observed scores, regression weighted scores, and multilevel weighted scores.

The results showed similar performance for the regression and multilevel weighted score methods. The observed score and weighted score methods performed similarly for test lengths of 20 and 40 items. The weighted score methods demonstrated advantages over the observed score method for test lengths of 5 and 10 items, but the improvement was small and inconsistent. Techniques for improving the performance of the weighted score methods are discussed.

[1] Paper presented at the Annual Meeting of the American Educational Research Association, April 18-22, 1995, San Francisco, CA.

Demonstrating the Utility of a Multilevel Model in the Assessment of Differential Item Functioning

The Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988) is a popular method for examining differential item functioning (DIF) in comparison groups. The method is designed to detect DIF in dichotomously scored items, for groups that are matched on some measure of ability; an item is said to display DIF if members of the two groups with the same ability have different probabilities of answering the item correctly. Routinely, total test score is employed as the matching criterion for the MH, due to its ready availability and the assumption that observed score is representative of the latent ability or true score of the examinee. With multiple choice tests, test length is usually sufficient so that the MH is not distorted by an unreliable observed score; with short tests, observed score becomes increasingly unreliable, and the MH statistic becomes unstable (see Donoghue & Allen, 1993). Donoghue, Holland, and Thayer (1993) concluded that test lengths of fewer than 20 items should not be used to construct matching variables.

With cognitive testing moving toward the use of performance assessment tasks, the issue of matching becomes more pertinent, as the number of items is typically very small and observed score is unreliable. A new matching technique for the MH procedure is required if the method is to be used to assess short tests with dichotomously scored responses or extended for the detection of DIF in polytomously scored assessments.

Research on External and Internal Matching Criteria for DIF Studies

Evaluations of alternatives to matching on observed score have been limited primarily to the assessment of internal criteria; few studies have been published where external criteria were used in matching the comparison groups. Zwick and Ercikan (1989) hypothesized that much of the observed DIF on the NAEP History assessment was a result of group differences in history coursework, and that the occurrence of DIF would be reduced if the groups were more similar in coursework. They conditioned on the number of historical periods studied within each level of observed score, but found the number of MH DIF items was not reduced when compared to DIF items produced by matching on observed score. In an extension of the MH procedure to assess DIF in polytomously scored items, Zwick, Donoghue, and Grima (1993) matched on an observed score that summed 20 dichotomous items and five polychotomous items. The authors addressed the need for research on aspects of matching such as multidimensional matching variables, measures external to the assessment, and matching variables based on noncognitive data.

Multiple studies (Clauser, Mazor, & Hambleton, 1991; Clauser, Nungester, Mazor, & Ripkey, 1994; Mazor, Narayanan, Stout, & Roussos, 1994; and Tian, Pang, & Boss, 1994) have reported that the choice of matching criterion, total test or a subtest, resulted in differential classification of DIF in items. In contrast, Ryan (1991) studied the effect of the number and context of test items comprising total test score and found no clear tendency for items displaying a certain magnitude of DIF to vary across matching criteria.

Additional studies have assessed the effect of collapsing score categories on the performance of the MH statistic. Raju, Bode, and Larsen (1989) concluded that 4 or more score groups of equal width were sufficient to yield stable MH estimates on a 40-item test. Clauser, Mazor, and Hambleton (1994) found a gain in power associated with reducing the number of score groups for an 80-item test. Donoghue and Allen (1993) showed that some methods of matching on collapsed score categories (defined as thick matching) improved the performance of the MH procedure for short tests, while matching on total test score (defined as thin matching) yielded the best results for long tests. Despite the improvement in the performance of the MH procedure implied by methods of thick matching in these studies, caution must be employed in collapsing score categories because the procedure increases the possibility of confounding DIF with a difference in the average ability of the groups (often called impact) as the number of score categories decreases.

The Research Problem

This research examined the effect of an alternate matching strategy on the performance of the MH for dichotomously scored items, seeking an alternative to matching on observed score that would yield more stable results under short test lengths where the MH matched on observed score performs poorly. External background variables were incorporated into the matching criterion with observed score in an attempt to match on an estimate closer to true score than observed score. Matching on the alternate estimate was expected to result in more reliable MH estimates where the MH is known to be unstable.

The alternate matching used a weighted score that was a function of the observed score and a predicted score derived from background variables external to the test. The predicted score was obtained using either a regression model or a multilevel model. In the regression model, observed score was regressed on student-level background variables to obtain the predicted score. In the multilevel model, the predicted score was computed by modeling student characteristics within school district characteristics.

Given the predicted score (either from the regression or the multilevel model), a weighted score was computed as the sum of the observed and predicted scores, each weighted by the inverse of their error variance, divided by the sum of the weights. The weighted score comprised both an internal criterion (observed score) and an externally-based criterion (predicted score). Of interest was the extent to which the competing methods of matching resulted in the correct identification of the direction and magnitude of DIF within the same items, whether the alternate matching method correctly identified DIF where the MH matched on observed score failed, and the relative power of the two methods when the presence of DIF was correctly identified.

The Standard Mantel-Haenszel Procedure

The MH procedure was developed by Mantel and Haenszel (1959) for cancer research; the applicability of the MH procedure to the detection of DIF for two comparison groups was demonstrated by Holland and Thayer (1988). Under this approach, the performance of a focal group on an item of interest (the studied item) is compared to the performance of a reference group, where the reference group provides a standard for comparison. For a studied item, each observed score category s is represented with a 2 × 2 table of group by item response, where $s = 0, 1, ..., k$ for a k-item test:

	Correct	Incorrect	Total
Reference	R_R	W_R	N_R
Focal	R_F	W_F	N_F
Total	R_s	W_s	N_s

The table cells contain the frequencies of correct responses to the item in the reference group (R_R), the focal group (R_F), and the combined group (R_s) at s; the frequencies of incorrect responses to the item in each group at s (W_R, W_F, and W_s); and the total number of examinees within each group at s (N_R, N_F, and N_s).

The MH common-odds ratio estimator is computed for each item as

$$MH = \frac{\sum_{s=0}^{k} \frac{R_R W_F}{N_s}}{\sum_{s=0}^{k} \frac{R_F W_R}{N_s}} . \tag{1}$$

When MH = 1, the null hypothesis of equal odds of a correct response between comparable members of the reference and focal groups is met:

$$H_0: \frac{R_R}{W_R} = \frac{R_F}{W_F} . \tag{2}$$

When MH ≠ 1, the alternative hypothesis holds:

$$H_1: \frac{R_R}{W_R} = MH \frac{R_F}{W_F} . \tag{3}$$

The MH value will be greater than 1.0 when the item favors the reference group and less than 1.0 when the item favors the focal group. The hypothesis may be evaluated using a one degree of freedom chi-square test that is the uniformly most powerful unbiased test of the null hypothesis (see Holland & Thayer, 1988). The chi-square test has the form

$$MH\,\chi^2 = \frac{\left(\left| \sum_{s=0}^{k} R_R - \sum_{s=0}^{k} E(R_R) \right| - \frac{1}{2} \right)^2}{\sum_{s=0}^{k} Var(R_R)} , \tag{4}$$

where

$$E(R_R) = \frac{N_R R_s}{N_s} \tag{5}$$

and

$$Var(R_R) = \frac{N_R N_F R_s W_s}{N_s^2 (N_s - 1)} . \tag{6}$$

The Multilevel Model for the Research Problem

Data that fall into a hierarchical structure where one level of measurement is nested within another may be modeled using multilevel techniques. The data used in this study were of the form students nested within school districts. The multilevel model corresponding to this structure is a two-level model, where students are represented at Level-1 (the within-school districts model) and the school districts are represented at Level-2 (the between-school districts model). The Level-1 model measures variation within school districts:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \ldots + \beta_{pj}X_{pij} + \epsilon_{ij} \tag{7}$$

for $j = 1,2, \ldots, n$ school districts. X_{pij} is a measured characteristic of student i within school district j, β_{pj} represents the expected change in the outcome for a fixed unit of X_p for student i within school district j, and ε_{ij} is random error. The error structure may vary, but commonly it is assumed to be normal with a mean of zero and constant variance, σ^2. Both the individual outcomes, Y_{ij}, and the regression parameters, β_{pj}, are assumed normally distributed.

In the multilevel model, each parameter represented in the Level-1 model is allowed to vary across the units of analysis at Level-2. This variation is represented in an equation modeling the intercepts and slopes as outcomes at the second level. The Level-2 model is a between-school districts model:

$$\beta_{pj} = \gamma_{p0} + \gamma_{p1}Z_{1j} + \gamma_{p2}Z_{2j} + \ldots + \gamma_{pq}Z_{qj} + r_{pj} \tag{8}$$

where Z_{qj} is a measured characteristic of the school district, γ_{pq} represents the effect of Z_{qj} on the *pth* parameter for school district j, and r_{pj} is random error. The error structure at Level-2 is assumed to be multivariate normal, with mean zero and covariance matrix T. The parameters of the between-school districts model are fixed effects, while the errors are random effects of the measured characteristics that influence the parameters associated with each school district.

Parameter estimates can be obtained for the fixed effects, the random Level-1 coefficients, and the variance-covariance components of the model using the computer program HLM (Bryk, Raudenbush, Seltzer, & Congdon, 1989). The estimates of the first level parameters are said to be optimal in that no other estimates have a smaller expected mean-squared error. The estimates are often called shrinkage estimates because the ordinary least squares regression line for each school district is pulled toward a predicted value based on the school district-level model. When no second level predictors are included in the model, shrinkage is toward a common regression line. The amount of shrinkage is conditional on the reliability of the first level parameters; when the first level parameters are reliable, shrinkage is minimal.

These estimates have been shown empirically to be more stable and less variable in time than least squares estimates, and predict better in the future (Rubin, 1980). Application of a multilevel model to predicting first year performance in law school avoided the bouncing beta problem, describing the fact that the estimates fluctuate wildly from year to year when alternate samples are employed (Rubin, 1989). The estimates also perform better in cross-validation than the classical estimates (Braun, 1989; Thissen & Bock, 1990).

When the data are of a hierarchical nature, the multilevel method is preferable to regression or other prediction techniques because it allows unique information for each school district to be included in the prediction model. However, using regression techniques to obtain the predicted score may be more desirable, due to the greater complexity of the task of obtaining a predicted score under a multilevel model. If a comparison of the MH estimates based on the multilevel and regression predicted scores shows little difference across methods of obtaining predicted score, then regression would be the preferred method for computing the predicted score in future applications.

Method

The research involved two stages. In the first stage, multilevel and regression models for the predicted score were determined using observed score on the 80-item Grade 3 North Carolina End of Grade Mathematics Test and available student and school district data obtained from the North Carolina Department of Public Instruction. Three alternate forms of the Mathematics test were administered in the spring of 1993 in North Carolina's public schools, with approximately 25,000 students responding to each form[2]. The second stage comprised a simulation study in which the models from the first stage were used to generate data, and MH statistics were then computed based on observed score and weighted score matching. Test parameters and item response data were simulated in this study to allow

[2] Preliminary analyses showed responses to all forms of the Grade 3 Mathematics Test to be unidimensional.

experimental control of the occurrence, amount, and direction of DIF in an item, enabling direct comparisons to be made of results from each matching criterion.

The Prediction Models

The Multilevel Model

The predictors for the multilevel model used in the study were chosen from the available student-level and school district-level variables to yield the most parsimonious, best-fitting model. The model was fittcd to a 10% sample (N = 2271) of respondents to Form A of the Grade 3 Mathematics test and cross-validated on a 10% sample (N = 2191) of respondents to Form B of the test. The final multilevel model included the Level-1 model

$$\mathbf{MSCORE} = \beta_0 + \beta_1(\mathbf{PEDUC}) + \beta_2(\mathbf{EXCEPT}) + \epsilon \qquad (9)$$

and the Level-2 models

$$\beta_0 = \gamma_{00} + \gamma_{01}(\mathbf{LOCALEXP}) + r \qquad (10)$$

$$\beta_1 = \gamma_{10} + \gamma_{11}(\mathbf{LOCALEXP}) + r \qquad (11)$$

$$\beta_2 = \gamma_{20} + \gamma_{21}(\mathbf{LOCALEXP}) + r\,. \qquad (12)$$

In the model, MSCORE represented the observed score on the 80-item test, PEDUC represented a teacher-report of the highest educational level attained by either of the student's parents (1 = did not finish high school, 2 = high school graduate, 3 = trade/business school, 4 = community college/technical college/private junior college, 5 = four-year college, and 6 = graduate school), EXCEPT represented a teacher-classification of the exceptional status of the student (0 = gifted, 1 = not exceptional, and 2 = exceptional[3]), and LOCALEXP represented the local expenditure of the school district per student[4].

[3] Students classified as exceptional comprised those who were multihandicapped, autistic, behaviorally-emotionally handicapped, deaf-blind, hearing impaired, mentally handicapped, orthopedically impaired, specific learning disability, speech-language impaired, visually impaired, or had some other health impairment.

[4] According to the North Carolina Department of Public Instruction, all expenditures not funded by the state or the federal governments are regarded as "local" expenditures. As such, they include all funds supplied from local governments and other local sources.

The Regression Model

While the multilevel model provides information about the school districts at the second level, additional information beyond the student level may be unnecessary in this context. For purposes of comparison, a second predicted score was computed, based on a regression model that regressed observed score on the predictors from the student level of the multilevel model:

$$\mathbf{MSCORE} = \beta_0 + \beta_1(\mathbf{PEDUC}) + \beta_2(\mathbf{EXCEPT}) + \epsilon \,. \tag{13}$$

The use of this model allowed the school district-level information in the multilevel prediction model to be assessed. Operating at the student-level only ignored the nesting of students within school districts.

The Simulation Study

Item responses and the background variables for examinees were simulated and MH statistics were computed using a computer program written in Fortran[5].

Fixed Conditions

The inclusion of the studied variable in the matching criterion, sample size, the ratio of sample sizes in the comparison groups, the degree of incongruence (or lack of overlap) between the comparison distributions, and the number of DIF items in the matching criterion were fixed conditions in the study. Given the results of research regarding the inclusion or exclusion of the studied variable from the matching criterion (Donoghue, et al., 1993; Holland & Thayer, 1988; Zwick, 1990), the studied variable was included in the matching criterion in the simulation. The sample size was fixed at ($N_R + N_F = 1000$), with the sample size for the focal group fixed at 30% of the total sample size ($N_F = 300$ and $N_R = 700$). The proportion of focal group members was set at the proportion of blacks observed in the population of blacks and whites on the Grade 3 Mathematics test. The reference group was drawn from a $N(0,1)$ distribution, while the focal group was drawn from a $N(-1, 1)$ distribution. This difference between the two distributions is slightly larger than that observed in the empirical data. Finally, DIF was induced only in the studied item, so that the MH was not distorted by DIF in the remaining items. When the studied item contained no DIF, there were no DIF items in the matching criterion.

[5] The author wishes to thank Judy Spray for providing portions of the Fortran code.

Manipulated Conditions

The conditions that were manipulated in this study were test length, degree of DIF in the studied item, item difficulty, and item discrimination. The resulting design for the study was a 4 × 2 × 2 × 2 factorial design, with 4 levels of test length (5, 10, 20, and 40), 2 levels of DIF (0, and 0.6), 2 levels of the *b* parameter (-0.5 and 0.5), and 2 levels of the *a* parameter (0.5 and 1.0). This design yielded a total of 32 research conditions that were repeated over each of three matching criteria (observed score, regression weighted score, and multilevel weighted score).

Test length

In this study, four levels of test length were used: 5, 10, 20, and 40 items. Monte Carlo studies have shown test lengths of 20 and 40 items to yield satisfactory results, and short tests (5-10 items) to yield unsatisfactory results (see Donoghue & Allen, 1993; Donoghue, et al., 1993). Of primary interest was the functioning of two weighted score MH estimates (based on predicted scores from regression and multilevel models) on short tests, where the observed score MH has been found to perform poorly.

Degree of DIF in the Studied Item

DIF was induced in the studied item by manipulating the difficulty parameter of a unidimensional three-parameter logistic item response function:

$$P(\theta) = c + \frac{(1-c)}{1 + e^{-1.7a(\theta-b_i)}} . \qquad (14)$$

In the "Non-DIF" condition of the study, the item parameters were the same across both the focal and reference groups (i.e., $a_F = a_R$, $b_F = b_R$, and $c_F = c_R$). In the "DIF" condition, DIF was induced in the difficulty parameter by setting $b_F = b_R + 0.6$. Inducing DIF in the difficulty parameter yields nonuniform, unidirectional DIF (see Cressie & Holland, 1983); this method is typically used in DIF studies.

Item Difficulty, Item Discrimination, and Guessing Parameters

In order to minimize the confounding of the MH value with item difficulty, the levels of the difficulty parameter for the studied item were set to the moderate values of -0.5 and 0.5. The levels of the item discrimination parameter for the studied item were set to 0.5 and 1.0, and a constant value of $c_R = c_F = 0.25$ was used for the guessing parameter. The guessing parameter represented the chance probability of answering correctly on the 4-alternative format of the Grade 3 Mathematics test.

For all items within a test excluding the studied item (i.e., the core items), the parameter values were randomly selected. The *a* parameters were randomly drawn from a *N*(1,0.25) distribution with values greater than 1.5 truncated to 1.5, and values less than 0.5 truncated to 0.5. The *b* parameters were randomly drawn from a *N*(0,0.75) distribution with extreme values truncated to -1.5 and 1.5. The *c* parameters were uniformly distributed between 0.05 and 0.25. While the parameter values varied across the core items in a test, the parameters of the core items were fixed across the focal and reference groups, so that no DIF occurred in those items.

Computatic.ı of Weighted Score

In each analysis condition, wo separate weighted scores were computed for each examinee as the sum of observed score (X) weighted by its inverse error variance and a predicted score (PS) weighted by its inverse error variance, divided by the sum of the weights:

$$WS = \frac{\frac{1}{\sigma^2_{e_X}}(X) + \frac{1}{\sigma^2_{e_{PS}}}(PS)}{\frac{1}{\sigma^2_{e_X}} + \frac{1}{\sigma^2_{e_{PS}}}}, \qquad (15)$$

with

$$\sigma^2_{e_X} = \sigma^2_X(1 - \rho_X) \qquad (16)$$

and

$$\sigma^2_{e_{PS}} = \frac{\Sigma(X-PS)^2}{df}. \qquad (17)$$

Observed score was computed as the sum of the item responses. Coefficient alpha was used as the estimate of test reliability (ρ_X). For the first weighted score, predicted score was computed from fitting the regression model in Equation 13 to the item response data generated in that condition. For the second weighted score, predicted score was computed from fitting the multilevel model in Equations 9-12. Each weighted score was then rescaled to a 40-item scale for all test lengths. Recent findings for the performance of the MH statistic with collapsed numbers of score categories for samples with unequal ability distributions indicated the suitability of the rescaling (Donoghue & Allen, 1993; Clauser, Mazor, & Hambleton, 1994).

MH and χ^2 Computation

For each score category of observed score and the two weighted scores, a 2 × 2 table of frequencies for group by item response was created. Three MH estimates were computed for each item: a MH based on observed score matching, a MH based on weighted score matching where predicted score was derived from the regression model, and a MH based on weighted score matching where predicted score was derived from the multilevel model. In addition, MH χ^2 statistics were computed for each estimate to provide a standard for the magnitude of the estimated DIF, and allow comparison of the performance of the MH across the three methods of matching. The MH estimates and χ^2 statistics were replicated 200 times in each experimental condition.

Results

Comparison of Mean Performance

In order to statistically evaluate the differences in performance of the three methods of matching within the Non-DIF and DIF conditions, repeated measures analysis of variance tests with Helmert contrasts were performed. The Helmert contrasts tested the average observed score against the average of the weighted scores and the average regression weighted score against the average multilevel weighted score. In the Non-DIF condition, the square roots of the MH χ^2 values were analyzed because the magnitude of the effect was of interest, but not the direction[6]. In the DIF condition, the log of the MH odds-ratios (LMH) were analyzed because the direction and the magnitude of the effect were of interest[7].

[6] The analysis was performed on the square roots of the χ^2 values to avoid problems of interpretation due to the skewness of the χ^2 distribution.

[7] The distribution of MH odds-ratios ranges from 0 to positive infinity. The transformation of the MH odds-ratio to a log odds-ratio shifts the distribution from negative to positive infinity. A MH value less than one corresponds to a negative log odds-ratio, while a MH value greater than 1.0 corresponds to a positive log odds-ratio.

Ncn DIF Condition

The χ^2 means and standard deviations for the Non-DIF condition are reported in Table 1. The !6 analysis conditions represented are for the levels of test length (40, 20, 10, and 5) crossed with the levels of the *a* parameter (.5, 1) and the levels of the *b* parameter for the studied item (-.5, .5). In all tables, the levels of the *b* parameter represent the value of the studied item in the reference group. The χ^2 averages and standard deviations collapsed across conditions of *a* and *b* within test length are presented in Table 2. The results of the Helmert contrasts for the Non-DIF condition are presented in Table 3. The test of the regression weighted score method against the multilevel weighted score method showed no significant effects. Because the analysis showed no differences between the two weighted scores, the weighted score methods will be considered together for conditions where no DIF was induced in the studied item.

For the test of the observed score mean versus the average of the weighted score means, a significant interaction ($\alpha = .05$) was found among test length, the item discrimination parameter (*a*), and the item difficulty parameter (*b*). The observed score method performed slightly better than the weighted score methods at test lengths of 5, 10, and 20 items where $a = 1$ and $b = .5$, and at test lengths of 10 and 20 items where $a = .5$ and $b = -.5$. In all other conditions, the weighted score methods performed better than the observed score method, with the largest difference occurring at a test length of 5 items. At test lengths of 20 and 40 items, the difference was very similar across the conditions of the *a* and *b* parameters. The weighted score methods showed the most improvement over the observed score method in the condition where $a = .5$ and $b = .5$, for test lengths of 5 and 10 items.

A significant interaction was also found between the levels of the *a* and *b* parameters for the contrast of observed score versus the weighted score methods. When the studied item was more difficult ($b = .5$), the weighted score methods performed better than the observed score method for the less discriminating item ($a = .5$) and slightly worse than the observed score method for the more discriminating item ($a = 1$). When the studied item was less difficult ($b = -.5$), the weighted score methods performed better than the observed score method, with an increase in the difference as the studied item moved from less discriminating ($a = .5$) to more discriminating ($a = 1$). Overall, the difference in performance between the observed and weighted score methods was largest where $a = .5$ and $b = .5$ and smallest where $a = 1$ and $b = .5$.

The contrast of observed score versus the weighted score methods also showed a significant effect for test length, and a significant effect for overall mean. The significant effect for test length showed a better performance for the weighted score methods than the observed score method at all test lengths, with the largest differences in performance occurring for test lengths of 5 items. The significant effect for the mean showed a lower average χ^2 over all analysis conditions for the weighted score methods.

DIF Condition

The MH means and standard deviations for the DIF condition are presented in Table 4. The MH averages and standard deviations collapsed across conditions of a and b within test length are presented in Table 5. The results of the Helmert contrasts for the DIF condition are presented in Table 6. The test of the regression weighted score model against the multilevel weighted score model showed a significant effect for test length and for the average over all conditions. At all test lengths, the multilevel weighted score method performed better (i.e., had a higher average MH value) than the regression method, although the differences were very small. The comparison of MH values, averaged over all analysis conditions, showed a better performance for the multilevel weighted score method overall. Because the effect for test length was very small and no other effects were significant, the multilevel weighted score and the regression weighted score methods will be considered together for conditions where DIF was induced in the studied item.

For the test of observed score MH versus the average MH of the weighted score methods, significant interactions were found between the item discrimination parameter (a) and the item difficulty parameter (b), test length and the item difficulty parameter, and test length and the item discrimination parameter. Significant main effects were found for the item difficulty parameter, the item discrimination parameter and the overall mean. Unlike the Non-DIF condition, a significant effect was not found for the interaction between test length, item discrimination, and item difficulty.

The interaction between test length and item difficulty showed that the observed score method performed better than the weighted score method at all test lengths when the item was less difficult ($b = -.5$), with the difference in performance decreasinp as test length increased.

The observed score method also performed slightly better at 20 and 40-item test lengths when *b* = .5. The weighted score method performed slightly better than the observed score method at test lengths of 5 and 10 items where the item was more difficult (*b* = .5) with a decrease in the difference in performance as test length increased from 5 to 10 items. The differences in performance at each level of the *b* parameter were very small at test lengths of 20 and 40 items.

The interaction between test length and item discrimination showed that the observed score method performed better than the weighted score method at all test lengths when the item was more discriminating (*a* = 1), with the difference in performance decreasing as test length increased. The observed score method also performed slightly better at 20 and 40-item test lengths when *a* = .5. The weighted score method performed slightly better than the observed score method at test lengths of 5 and 10 items where the item was less discriminating (*a* = .5). The differences in performance at each level of *a* were very small at test lengths of 20 and 40 items. The trends observed across test length for the significant interactions show a greater effect for the *a* parameter than the *b* parameter on the performance of the matching methods.

Although the observed score method offered an advantage at all levels of the *a* and *b* parameters on test lengths of 20 and 40 items, the advantage appears to be slight. The results indicate that the best performance of the weighted score methods occurred at test lengths of 5 and 10 items where the studied item was less discriminating and more difficult (*a* = .5, *b* = .5). The best performance for the observed score method occurred when the studied item was more discriminating and less difficult (*a* = 1, *b* = -.5). The advantage offered by the observed score method over the weighted score methods in the remaining two conditions appeared to be very slight.

For the significant main effects, the analysis showed a superior performance of the observed score method where the studied item was more discriminating (*a* = 1) and where the studied item was less difficult (*b* = -.5). The weighted score methods performed better than the observed score method when the studied item was less discriminating (*a* = .5) and where the studied item was more difficult (*b* = .5). The effect for the mean showed a higher MH average over all analysis conditions for the observed score method than for the weighted score methods.

Hit Rates and False Positive Identifications

The hit rates and false positive identifications of DIF provide information regarding the performance of the MH statistics across repeated applications of the DIF detection procedure. The hit rates and false positive rates are presented as the percentage of times the studied item produced a χ^2 statistic at the .05 level of significance, across 200 replications of the analysis condition.

False Positives

The false positives for the studied item in which no DIF was induced are presented in Table 7. Within a test length of 40 items, the observed score method and the weighted score methods performed consistently, across the four combinations of the *a* and *b* parameters. The results were only slightly less consistent for 20-item tests. For a test length of 10 items, the hit rates were still fairly consistent across the three methods of matching; the results of the analysis of variance of the chi-squares suggest that the superior performance of the weighted score methods when $a = .5$ and $b = .5$ is significant, while the other differences may represent chance fluctuations.

When test length was reduced to 5 items, the differences in false positive rates for the observed score matching method and the two weighted score matching methods increased. In the condition where the studied item was more discriminating and more difficult ($a = 1$, $b = .5$), the three methods were consistent in the number of items flagged. The differences in the percentage of false positives identified by the observed score method and the weighted score methods were more inconsistent across the remaining conditions. The results of the interaction between test length, item discrimination, and item difficulty for the analysis of variance of the chi-squares indicate that the largest differences in performance between the observed score method and the weighted score methods occurred at test lengths of 5 items.

Hit Rates

The hit rates for the studied item in which DIF was induced are presented in Table 8. For a test length of 40 items, the three methods of matching performed consistently, across the four combinations of the *a* and *b* parameters. For 20-item tests, the results were only slightly less consistent across the four combinations of the *a* and *b* parameters. For a test length of 10 items, the hit rates were less consistent across the three methods of matching for the condition where the studied item was more discriminating and less difficult ($a = 1$, $b =$

-.5). In the other three conditions of the *a* and *b* parameters, the three methods were more consistent in the amount of items identified. When test length was reduced to 5 items, the differences between observed scoie matching and the two weighted score matching methods remained closest in the two conditions where the studied item was more difficult (*b* = .5).

In the conditions where the studied item was less difficult (*b* = -.5), the hit rates across test length were slightly higher for the observed score method than the weighted score methods, with the difference in performance decreasing as test length increased. However, the results of the analysis of variance suggest that at test lengths of 5 and 10 items, the hit rates should be slightly better for the weighted score methods than the observed score method where the studied item is more difficult (*b* = .5); the difference in hit rates appears to be due to the differences among the variances of the statistic. The results also suggest that for test lengths of 20 and 40 items, the hit rates should be fairly consistent across the levels of *a* and *b*. At test lengths of 5 and 10 items, the hit rates should be slightly better for the weighted score methods than the observed score method where the studied item is less discriminating (*a* = .5).

Performance Across DIF and Non-DIF Conditions

A comparison of the hit rates and false positive identifications of DIF in Tables 7 and 8 shows that the overall performance of the three methods of matching was best where the studied item was more discriminating and less difficult (*a* = 1, *b* = -.5). This finding corresponds to previous conclusions that DIF will be most easily detected in items of moderate difficulty and high discrimination (Rogers & Swaminathan, 1993). For the DIF-induced items within this condition, the reduction in hit rates across matching methods was very slight as test length decreased from 40 to 5 items. For the Non-DIF items, the number of false positive identifications of DIF for the matching methods remained fairly constant across test lengths of 40, 20, and 10 items, with an increase in the rate of false positives observed for each method at a test length of 5 items.

The overall performance of the three matching methods was the poorest where the studied item was less discriminating and more difficult (*a* = .5, *b* = .5). In this condition, the total number of false positive identifications of DIF exceeded the total number of correct identifications of DIF for the observed score method of matching. The two weighted score methods performed only slightly better, identifying a few more DIF-induced items than Non-

DIF items. This discrepancy between the hit rates and false positives is largely attributable to the performance of the methods for test lengths of 10 and 5 items. When test length was 40 or 20 items, all three matching methods were consistent in their rate of flagging for both the DIF and Non-DIF conditions with substantially higher hit rates than false positive identifications. Reducing the number of items in the matching criterion below 20 appears to increase the likelihood that DIF will go undetected in an item. A similar trend was noted by Donoghue and Allen (1993), who found a tendency for 5 and 10-item tests to falsely identify Non-DIF items as possessing DIF.

For test lengths of 40 and 20 items within the Non-DIF and DIF conditions, it appears that neither weighted score method offers an improvement over the observed score method. Inconsistencies in the hit rates and false positives for the three methods did not appear until test length was decreased to 10 and 5 items. Any potential improvement in the performance of the MH by the weighted score matching technique is most noticeable in the case of 5 items in which no DIF is induced, particularly where the studied item was less discriminating and more difficult (a = .5, b = .5).

Summary

The hit rates, false positive rates, χ^2 averages, and MH averages generally displayed similar performances between the three methods of matching for test lengths of 20 and 40 items. A difference in the performance between the three methods of matching became more distinct in the case of 5 and 10-item test lengths. Overall, the results suggested a better performance of the weighted score methods, relative to the observed score method when the studied item was less discriminating (a = .5), with a better performance for the observed score matching method when the studied item was more discriminating (a = 1). Likewise, the weighted score methods seemed to perform better when the studied item was more difficult (b = .5), while the observed score method appeared to perform better when the studied item was less difficult (b = -.5). The two weighted score methods showed similar performances across the DIF and Non-DIF conditions.

The combination of a more discriminating and less difficult studied item (a = 1, b = -.5) showed the poorest performance of the weighted score methods, relative to the observed score method. The detection of DIF was best for all three methods under this combination of the a and b parameters because the item discriminated between individuals midway between the means of the comparison group distributions. The combination of a less discriminating and more difficult studied item (a = .5, b = .5) showed the best performance of the weighted score methods, relative to the observed score method. Overall, the detection of DIF was poorest within this condition; each method showed a greater percentage of false positives than hits on test lengths of 5 and 10 items. The MH statistic performed poorly because the item was difficult for the focal group and did not discriminate well between individuals with differing levels of ability.

For the 5 and 10 item test lengths in the condition a = .5, b = .5, and the 5-item test lengths where a = .5, b = -.5 and a = 1, b = .5, the MH log odds-ratio is biased, showing DIF in favor of the focal group (i.e., a negative value) when no DIF was induced. A similar bias was noted by Donoghue and Allen (1993) on 5 and 10-item tests. The difference between the focal and reference group distributions may be confounded with DIF for short tests in these conditions. When DIF was induced against the focal group the log odds-ratios shifted right (i.e., became less negative); however, because of the bias in the statistic, the estimates remained close to zero, and the hit rates were very low.

For these conditions of the study, the bias in the MH statistic resulted in a greater percentage of false positives than correct identifications of DIF. Although the weighted score methods showed a slight reduction in bias for the MH statistic, they still resulted in more false positives than correct identifications of DIF with the shorter test lengths. Given this discrepancy in performance, employing the weighted scores of this study as the matching criterion did not provide a particularly useful alternative to the traditional method of matching on observed score, or enable the use of the MH procedure with short tests.

Discussion

Although the weighted score MH estimates did not perform consistently better than the observed score MH estimates in the simulation study, the weighted scores correlated more highly with latent ability (θ) than did the observed score. It is likely that there was little improvement in the performance of the MH statistic with weighted score matching because the correlation between the weighted score and θ was only slightly larger than the correlation between the observed score and θ. As test length decreased the magnitude of the difference between the correlations of the weighted scores with θ and the correlation of the observed score with θ increased, which suggests an explanation for the best performance of the weighted score methods at test lengths of 5 and 10 items. However, the observed correlations suggest that the strength of the relationship between the weighted score and θ was not of a magnitude to enable the weighted score matching technique to perform optimally. For a MH statistic matched on a weighted score to perform well on shorter tests, the correlation between the weighted score and θ should be of a similar magnitude as the correlation between observed score and θ for test lengths of 20 or 40 items.

Effect of the Weighting Scheme

The observed trends in correlations across test length were probably a result of weighting the observed and predicted scores by the inverse of their respective error variances. At all test lengths, the error variance for the predicted score was larger than the error variance for the observeu score, with an increase in the difference between the error variances as test length increased. For test lengths of 20 and 40 items, the observed score contributed most to the weighted score; it is likely that the weighted score did not add enough information to improve noticeably upon the performance of the MH matched on observed score. At test lengths of 5 and 10 items, the observed score was much less influential and the background information contributed more to the weighted score, resulting in some improved performance of the weighted score MH.

Due to the greater cost of implementing the more complex procedure of weighted score matching, very little would be gained by employing weighted score matching at test lengths of 20 and 40 items, because observed score matching is effective. Improvement upon observed score matching is necessary in the case of 5 and 10-item tests, where the observed score method performs poorly. The results of the simulation study indicate that matching on a weighted score shows potential for shorter test lengths, but that better predictors would need to be used to show substantial gains.

Effect of Thin Matching

Although larger correlations were observed between the weighted score and θ than between the observed score and θ for all analysis conditions of the original simulation study, the results showed that the weighted score methods did not perform consistently better than the observed score. In the case of 20 and 40-item tests, this finding was not surprising because the increase in correlation between the weighted scores and θ over the correlation between the observed score and θ was very small. However, this occurrence was somewhat surprising in test lengths of 5 and 10 items, where the differences between the correlations of the weighted scores with θ and the correlations of the observed score with θ were larger. A possible explanatior for the inconsistent performance of the weighted score methods is that the rescaling of the weighted scores to a 40-item scale for all test lengths resulted in a loss of information for the MH matched on the weighted score. In rescaling to a 40-item scale it is probable that the 2 × 2 tables became more sparse for the shorter test lengths, eliminating some of the data. In the MH procedure, a score category with an empty row or column in the 2 × 2 table does not contribute to the computation of the odds-ratio, and the information from the remaining row or column is discarded. This would account for some of the inconsistencies in performance of the weighted score methods with the 5 and 10-item tests, and may also have contributed to the similar performance of the matching methods for test lengths of 20 and 40 items. Applying the weighted score method to a larger sample or to focal and reference group samples of equal size may help eliminate this problem.

A caution against employing overly fine matching was given by Donoghue and Allen (1993), who suggested that much of the data might be eliminated by such a process. They tested various methods of collapsing score categories as an alternative to finer matching. However, the classification onto a 40-item scale in this study allowed the unique background information for examinees to be retained in the matching process. Keeping the original scale for each test length would often have resulted in grouping together individuals with the same observed score but different background information, rather than distinguishing between them. As the study was intended to assess the contribution of the background information to the performance of the MH, the thinner matching was judged necessary for categorization.

Improving the Model

Additional simulations were conducted for test lengths of 5 and 10 items within the Non-DIF and DIF conditions. The purpose of the additional analyses was twofold: first, to study the effect of an improved prediction model on the performance of the weighted score methods, and second, to determine how much additional information was necessary to obtain the predicted score for the matching process. Improving the prediction model would decrease the error variance for the predicted score and allow the background information to contribute more to the weighted score. Greater contribution of the predicted score to the weighted score would result in a greater difference in the correlations between the weighted score and θ and the observed score and θ. Since using a multilevel model greatly complicates the task of computing a predicted score, a regression model would be preferable if it provides sufficient information for the matching procedure under a better predicting model.

In the new simulation, the performance of a third model was compared to the regression and multilevel models presented earlier. The new model is a regression model that includes the student-level predictors from the regression and multilevel models and the school-level predictor from the multilevel model, analyzed at the student level:

$$\mathbf{MSCORE} = \beta_0 + \beta_1(\mathbf{PEDUC}) + \beta_2(\mathbf{EXCEPT}) + \beta_3(\mathbf{LOCALEXP}) + \epsilon \,. \qquad (18)$$

This model allowed the degree of information necessary for an optimal performance of the weighted score methods to be addressed. The inclusion of this model examines whether the addition of another predictor (LOCALEXP) to the regression model appears sufficient for computation of the weighted score methods, or whether the more precise modeling of unique coefficients for each school appears necessary. Demonstrating the former would enable the use of the simpler regression technique for the matching process, while demonstrating the latter would require the use of the more complicated multilevel technique.

MH estimates and MH χ^2 statistics were computed over 100 replications for three values of model r^2 (.2778, .5, and .75). In the original simulation, r^2 was fixed at the value observed when the prediction model was fit to the real data (.2778). Within an analysis condition, the same seed was used to generate the data for the replications across the different r^2 values, making r^2 a within-unit factor. With each increase in the value of r^2, the error variance for the predicted score decreased, while the error variance for the observed score remained constant. Likewise, the correlations between the weighted scores and θ increased with each simulated improvement in model prediction, while the correlations between observed score and θ remained constant across the three values of r^2.

The false positive rates of identification for the Non-DIF condition are presented in Table 9, while the hit rates for the DIF condition are presented in Table 10. In the tables, the regression model containing the variables PEDUC and EXCEPT is labeled R1, while the regression model containing the variables PEDUC, EXCEPT, and LOCALEXP is labeled R2. For both test lengths in the condition a = 1, b = -.5, all methods appeared to perform fairly consistently across the levels of improved prediction. For test lengths of 5 items in the other conditions of a and b, all of the matching methods showed hit rates that were about the same or less than the false positive rates for an r^2 of .2778. As prediction improved, the regression weighted score including LOCALEXP and the multilevel weighted score showed more notable increases in hit rates and decreases in the number of false positives, while the identifications under the observed score method remained fairly constant. For test lengths of 10 items, the improvement in performance of these two methods was more distinct, with the multilevel weighted score showing the best performance. In the best predicting model, the weighted score methods appear to reduce the bias in the MH statistic for some conditions of a and b, moving from a greater percentage of false positives than correct identifications of DIF to a greater percentage of hits than false positives. That this trend did not occur in the 5-item case where $a = .5$, $b = .5$ suggests that an even better model would be required in this condition.

Additional information about the performance of the weighted score methods across the replications is presented in Tables 11 and 12. Table 11 gives the χ^2 averages and standard deviations for the Non-DIF condition, while Table 12 gives the MH averages and standard deviations for the DIF condition. For test lengths of 5 and 10 items, the χ^2 averages indicated a similar performance for the MH estimates matched on observed score across the values of r^2, while the χ^2 averages matched on the regression weighted score with LOCALEXP and multilevel weighted score showed decreasing averages as r^2 increased. This trend did not occur where $a = 1$, $b = -.5$, but in this condition all methods performed well, regardless of level of model prediction. Of the weighted score methods, the multilevel method showed the greatest decrease in the χ^2 averages and standard deviations as r^2 increased. The MH averages for the DIF condition showed corresponding trends across the values of r^2 for the methods of matching. Consistently, the observed score averages remained constant for both test lengths, while the weighted score averages increased with each increase in r^2. The biggest increases in average MH were noted in the case of the multilevel weighted score technique.

The averages and standard deviations, in addition to the hit rates and false positive rates, suggest the utility of a better prediction model for the weighted score matching technique. Also, as prediction improves, the difference in performance of the weighted score methods becomes greater. This suggests that the use of more unique information, such as is available with the multilevel model, is likely to lead to a more precise classification into score categories than occurs with using the regression models, resulting in an improved performance of the MH statistic. When prediction is as poor as observed in the real data, the distinction between the regression and multilevel models is small, and the performance of the weighted score methods does not differ.

Although the results suggest a better model will yield better results for matching on a weighted score, it may not be possible to find predictors that are not redundant with observed score that will result in the necessary level of prediction. However, a model with some degree of improved prediction, in conjunction with a weighting scheme that favors the observed score less, may increase the contribution of the predicted score to the weighted score, and yield a relationship between the weighted score and θ of the magnitude required for the MH to perform well.

Table 1. χ^2 averages and standard deviations across 200 replications for the Non-DIF condition ($\alpha = .05$), with Obs = observed score matching, Reg = regression weighted score matching, and ML = multilevel weighted score matching.

		a = .5, b = -.5			a = .5, b = .5		
Nitems		Obs	Reg	ML	Obs	Reg	ML
40	χ^2	.827	.818	.810	.928	.912	.876
	SD	1.288	1.243	1.240	1.350	1.308	1.245
20	χ^2	1.200	1.243	1.187	1.121	1.125	1.116
	SD	1.637	1.719	1.621	1.406	1.426	1.431
10	χ^2	1.447	1.472	1.436	1.811	1.734	1.641
	SD	2.249	2.210	2.202	1.993	1.931	1.816
5	χ^2	2.016	1.857	1.863	3.206	2.857	2.909
	SD	2.409	2.214	2.212	2.893	2.588	2.760
Total	χ^2	1.372	1.347	1.324	1.766	1.657	1.636
	SD	1.993	1.924	1.900	2.195	2.025	2.057

		a = 1, b = -.5			a = 1, b = .5		
Nitems		Obs	Reg	ML	Obs	Reg	ML
40	χ^2	.902	.898	.899	.876	.859	.838
	SD	1.478	1.507	1.493	1.343	1.337	1.292
20	χ^2	.800	.750	.802	1.236	1.271	1.202
	SD	1.255	1.192	1.284	1.819	1.952	1.799
10	χ^2	.956	.949	.921	1.476	1.561	1.467
	SD	1.282	1.284	1.216	1.990	2.097	2.052
5	χ^2	1.277	1.046	1.094	1.836	1.837	1.903
	SD	1.727	1.472	1.509	2.393	2.33	2.493
Total	χ^2	.984	.911	.929	1.356	1.382	1.352
	SD	1.456	1.371	1.383	1.952	1.994	1.993

Table 2. χ^2 averages and standard deviations across test length in the Non-DIF condition ($\alpha = .05$).

		Total		
Nitems		Obs	Reg	ML
40	χ^2	.883	.872	.856
	SD	1.364	1.350	1.320
20	χ^2	1.089	1.097	1.077
	SD	1.551	1.608	1.551
10	χ^2	1.423	1.429	1.366
	SD	1.933	1.933	1.876
5	χ^2	2.084	1.899	1.942
	SD	2.489	2.280	2.376
Total	χ^2	1.370	1.324	1.311
	SD	1.937	1.866	1.868

Table 3. Results of the repeated measures analysis of variance Helmert contrasts for the Non-DIF condition.

Contrast	Effect	F	P
Observed Score vs. Weighted Scores	Mean	30.98	.00
	Nitems	12.53	.00
	a	2.31	.13
	b	0.11	.73
	Nitems × a	2.52	.06
	Nitems × b	1.47	.22
	a × b	22.28	.00
	Nitems × a × b	8.17	.00
Regression vs. Multilevel	Mean	1.38	.24
	Nitems	2.30	.08
	a	2.43	.12
	b	2.53	.11
	Nitems × a	0.00	1.00
	Nitems × b	0.30	.83
	a × b	1.92	.17
	Nitems × a × b	0.51	.68

Table 4. MH averages and standard deviations across 200 replications for the DIF condition ($\alpha = .05$).

		a = .5, b = -.5			a = .5, b =..5		
Nitems		Obs	Reg	ML	Obs	Reg	ML
40	MH	1.389	1.385	1.388	1.258	1.256	1.262
	SD	.247	.247	.244	.214	.208	.210
20	MH	1.346	1.338	1.351	1.206	1.201	1.210
	SD	.208	.206	.213	.237	.241	.238
10	MH	1.280	1.278	1.288	1.109	1.108	1.116
	SD	.239	.246	.244	.166	.172	.168
5	MH	1.200	1.200	1.200	1.023	1.037	1.040
	SD	.212	.210	.204	.181	.190	.188
Total	MH	1.304	1.300	1.307	1.149	1.151	1.157
	SD	.238	.238	.237	.221	.220	.220

		a = 1, b = -.5			a = 1, b = .5		
Nitems		Obs	Reg	ML	Obs	Reg	ML
40	MH	1.808	1.799	1.809	1.342	1.336	1.339
	SD	.282	.290	.285	.238	.237	.234
20	MH	1.799	1.781	1.796	1.272	1.265	1.274
	SD	.334	.334	.340	.212	.207	.207
10	MH	1.819	1.780	1.799	1.211	1.205	1.212
	SD	.351	.340	.341	.205	.208	.207
5	MH	1.851	1.805	1.812	1.179	1.165	1.177
	SD	.353	.334	.341	.223	.209	.214
Total	MH	1.819	1.791	1.804	1.251	1.243	1.251
	SD	.331	.325	.327	.228	.225	.224

Table 5. MH averages and standard deviations across test length in the DIF condition (α = .05).

			Total	
Nitems		Obs	Reg	ML
40	MH	1.449	1.444	1.450
	SD	.325	.324	.324
20	MH	1.406	1.396	1.408
	SD	.343	.340	.343
10	MH	1.354	1.343	1.354
	SD	.371	.360	.362
5	MH	1.313	1.302	1.307
	SD	.405	.383	.385
Total	MH	1.381	1.371	1.380
	SD	.366	.356	.358

Table 6. Results of the repeated measures analysis of variance Helmert contrasts for the DIF condition.

Contrast	Effect	F	P
Observed Score vs. Weighted Scores	Mean	24.75	.00
	Nitems	0.64	.59
	a	77.24	.00
	b	36.43	.00
	Nitems × a	17.28	.00
	Nitems × b	8.38	.00
	a × b	4.23	.04
	Nitems × a × b	2.02	.11
Regression vs. Multilevel	Mean	97.58	.00
	Nitems	3.70	.01
	a	1.05	.30
	b	0.01	.93
	Nitems × a	1.04	.37
	Nitems × b	1.23	.30
	a × b	0.12	.73
	Nitems × a × b	0.67	.57

Table 7. Percentage of false positives across 200 replications in the Non-DIF condition (α = .05).

	a = .5, b = -.5			a = .5, b = .5		
Nitems	Obs	Reg	ML	Obs	Reg	ML
40	4	4	4	5.5	4.5	5.5
20	4	7.5	7	6.5	7	6.5
10	12	14.5	12.5	15.5	13	13.5
5	19	14	15.5	34	27	30
Total	10.75	10	9.75	15.38	12.88	13.88

	a = 1, b = -.5			a = 1, b = .5		
Nitems	Obs	Reg	ML	Obs	Reg	ML
40	3	3	3	3.5	3.5	3
20	5	3	4.5	7.5	8.5	8
10	4.5	4.5	4	9.5	10.5	10
5	9.5	6.5	6.5	16.5	15	16.5
Total	5.5	4.25	4.5	9.25	9.38	9.38

	Total		
Nitems	Obs	Reg	ML
40	4	3.75	3.88
20	6.75	6.5	6.5
10	10.38	10.63	10
5	19.75	15.63	17.13
Total	10.22	9.13	9.38

Table 8. Hit rates (as percents) across 200 replications in the Non-DIF condition (α = .05).

	a = .5, b = -.5			a = .5, b = .5		
Nitems	Obs	Reg	ML	Obs	Reg	ML
40	46.5	46	44.5	25	25	25
20	39.5	37	40	23	21.5	20.5
10	22.5	22	24.5	5	5.5	5
5	17.5	12.5	14.5	4	5	6
Total	31.5	29.38	30.88	14.38	14.25	14.13

	a = 1, b = -.5			a = 1, b = .5		
Nitems	Obs	Reg	ML	Obs	Reg	ML
40	94.5	94	94.5	35.5	33	36.5
20	88	89.5	88.5	22	22	21
10	90	84.5	88.5	15.5	16.5	16.5
5	89	86	88.5	12	10	11.5
Total	90.38	88.5	90	21.25	20.38	21.38

	Total		
Nitems	Obs	Reg	ML
40	50.38	49.5	50.13
20	43.13	42.5	42.5
10	33.25	32.13	33.63
5	30.75	28.38	30.13
Total	39.38	38.13	39.09

Table 9. False positives across 100 replications in the Non-DIF condition, for models with different values of *r*`.

			r^2 = .2778				r^2 = .5				r^2 = .75			
a	b	Nitems	Obs	R1	R2	ML	Obs	R1	R2	ML	Obs	R1	R2	ML
.5	-.5	10	9	12	8	10	7	8	5	8	13	4	5	5
		5	17	13	14	13	16	13	11	8	17	7	4	3
1	-.5	10	7	7	6	6	8	8	12	10	7	5	6	10
		5	9	6	7	8	8	7	8	7	10	9	8	13
.5	.5	10	14	9	12	11	13	10	6	7	14	9	5	6
		5	33	25	27	27	34	24	28	20	36	23	21	14
1	.5	10	7	10	11	8	6	9	8	7	10	7	3	5
		5	19	15	17	17	15	16	18	17	15	14	13	12

Table 10. Hit rates across 100 replications in the DIF condition, for models with different values of r^2.

			r^2 = .2778				r^2 = .5				r^2 = .75			
a	b	Nitems	Obs	R1	R2	ML	Obs	R1	R2	ML	Obs	R1	R2	ML
.5	-.5	10	22	20	23	25	25	28	33	33	26	40	43	48
		5	17	10	13	13	12	13	15	16	16	16	23	26
1	-.5	10	91	90	91	91	94	94	95	93	89	93	94	98
		5	89	87	88	88	89	88	90	88	86	87	91	94
.5	.5	10	5	6	8	5	8	9	11	7	7	10	13	19
		5	5	6	5	8	5	6	7	5	3	3	6	4
1	.5	10	11	12	12	14	10	12	16	17	9	17	25	29
		5	11	9	10	10	12	6	11	12	13	10	13	20

Table 11. χ^2 averages and standard deviations across 100 replications in the Non-DIF condition, for models with different values of r^2.

				$r^2 = .2778$				$r^2 = .5$				$r^2 = .75$			
a	b	Nitems		Obs	R1	R2	ML	Obs	R1	R2	ML	Obs	R1	R2	ML
.5	-.5	10	χ^2	1.202	1.257	1.189	1.208	1.166	1.062	.929	.947	1.315	.964	.923	.889
			sd	1.670	1.751	1.701	1.709	1.545	1.451	1.332	1.339	1.793	1.428	1.389	1.291
		5	χ^2	1.939	1.727	1.785	1.712	1.924	1.570	1.576	1.492	1.902	1.420	1.185	1.038
			sd	2.282	1.988	2.017	1.968	2.475	2.019	2.032	1.977	2.364	1.712	1.531	1.398
1	-.5	10	χ^2	1.019	1.006	1.050	.958	1.169	1.157	1.197	1.194	1.062	1.136	1.307	1.375
			sd	1.414	1.424	1.510	1.351	1.675	1.667	1.786	1.736	1.756	1.797	2.079	2.098
		5	χ^2	1.279	1.107	1.173	1.171	1.250	1.075	1.079	1.183	1.233	1.073	1.176	1.334
			sd	1.764	1.592	1.726	1.650	1.806	1.579	1.714	1.774	1.758	1.573	1.801	1.828
.5	.5	10	χ^2	1.613	1.632	1.612	1.583	1.558	1.422	1.270	1.176	1.647	1.179	1.023	1.014
			sd	2.047	2.199	2.197	2.093	1.969	1.989	1.749	1.701	2.010	1.671	1.394	1.480
		5	χ^2	3.177	2.901	2.905	2.906	3.220	2.686	2.686	2.433	3.264	2.389	2.119	1.740
			sd	2.585	2.388	2.422	2.463	2.751	2.348	2.270	2.173	2.898	2.315	2.189	1.847
1	.5	10	χ^2	1.391	1.550	1.510	1.447	1.289	1.326	1.174	1.124	1.360	1.101	.926	.893
			sd	1.805	2.113	2.068	2.011	1.896	1.948	1.888	1.771	2.046	1.815	1.651	1.459
		5	χ^2	1.897	1.866	1.892	1.890	1.797	1.747	1.699	1.635	1.972	1.873	1.490	1.352
			sd	2.473	2.442	2.394	2.540	2.309	2.246	2.218	2.205	2.512	2.424	1.970	1.911

Table 12. MH averages and standard deviations across 100 replications in the DIF condition, for models with different values or r^2.

				r^2 = .2778				r^2 = .5				r^2 = .75			
a	b	Nitems		Obs	R1	R2	ML	Obs	R1	R2	ML	Obs	R1	R2	ML
.5	-.5	10	χ^2	1.283	1.281	1.280	1.289	1.283	1.304	1.320	1.330	1.274	1.335	1.377	1.385
			sd	.237	.251	.242	.242	.224	.230	.234	.228	.224	.226	.233	.229
		5	χ^2	1.198	1.196	1.200	1.198	1.194	1.211	1.221	1.228	1.202	1.236	1.274	1.303
			sd	.210	.202	.207	.199	.207	.191	.201	.202	.213	.195	.202	.201
1	-.5	10	χ^2	1.838	1.809	1.809	1.825	1.836	1.838	1.861	1.877	1.830	1.893	1.942	1.976
			sd	.351	.341	.336	.342	.357	.351	.356	.365	.366	.366	.382	.378
		5	χ^2	1.838	1.788	1.785	1.804	1.850	1.804	1.823	1.856	1.835	1.818	1.892	1.932
			sd	.348	.323	.321	.340	.352	.335	.324	.350	.353	.340	.332	.341
5.	.5	10	χ^2	1.103	1.102	1.107	1.112	1.108	1.123	1.150	1.154	1.105	1.161	1.203	1.214
			sd	.167	.171	.172	.168	.168	.175	.173	.167	.169	.178	.171	.177
		5	χ^2	1.009	1.023	1.022	1.027	1.006	1.029	1.042	1.047	1.005	1.050	1.095	1.114
			sd	.179	.185	.184	.192	.175	.184	.184	.187	.159	.171	.174	.177
1	.5	10	χ^2	1.200	1.195	1.201	1.202	1.196	1.203	1.227	1.237	1.181	1.238	1.280	1.294
			sd	.182	.189	.186	.184	.178	.185	.182	.181	.168	.185	.190	.198
		5	χ^2	1.201	1.183	1.194	1.192	1.189	1.173	1.188	1.201	1.183	1.190	1.230	1.251
			sd	.199	.187	.193	.190	.198	.181	.182	.188	.205	.192	.198	.197

References

Braun, H.I. (1989). Empirical Bayes methods: A tool for exploratory data analysis. In R.D. Bock (Ed.) *Multilevel analysis of educational data* (Chapter 2, pp. 19-55). San Diego, Academic Press.

Bryk, A.S., Raudenbush, S.W., Seltzer, M., & Congdon, R.T. (1989). *An introduction to HLM: Computer program and user's guide* (2nd ed.). Chicago: University of Chicago Department of Education.

Clauser, B.E., Mazor. K., & Hambleton, R. (1991). Influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. *Applied Psychological Measurement, 15*, 353-359.

Clauser, B., Mazor, K.M., & Hambleton, R.K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31*, 67-78.

Clauser, B.E., Nungester, R.J., Mazor, K., & Ripkey, D. (1994, April). Detection of differential item functioning using the Mantel-Haenszel and logistic regression procedures. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Cressie, N., & Holland, P.W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika, 48*, 129-141.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18*, 131-154.

Donoghue, J.R., Holland, P.W., & Thayer, D.T. (1993). A Monte Carlo study of the factors that affect the Mantel-Haenszel and Standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (Chapter 7, pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel technique. In H. Wainer & H.I. Braun (Eds.), *Test validity* (Chapter 9, pp. 129-146). Hillsdale, NJ: Lawrence Erlbaum.

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.

Mazor, K.M., Narayanan, P., Stout, W., & Roussos, L. (1994, April). Identification of valid subtests for DIF analyses when tests are intentionally multidimensional. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Raju, N.S., Bode, R.K., & Larsen, V.S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. *Applied Measurement in Education, 2*, 1-13.

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

Rubin, D.B. (1980). Using empirical Bayes techniques in the Law School Validity Studies. *Journal of the American Statistical Association, 75*, 801-827.

Rubin, D.B. (1989). Some applications of multilevel models to educational data. In R.D. Bock (Ed.) *Multilevel analysis of educational data* (Chapter 1, pp. 1-17). San Diego, Academic Press.

Ryan, K. E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. *Journal of Educational Measurement, 28*, 325-337.

Thissen, D., & Bock, R.D. (1990). Linear and nonlinear curve fitting. In A. von Eye (Ed.) *Statistical methods in longitudinal research, Volume II* (pp. 289-318). San Diego, Academic Press.

Tian, F., Pang, X.L., & Boss, M.W. (1994, April). The effects of sample size and criterion variables on the identification of DIF by the Mantel-Haenszel and logistic regression procedures. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-197.

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

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55-66.