Educational and Psychological Measurement
Volume 68 Number 6
December 2008 940-958

10.1177/0013164408315265
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Generalized Mantel-Haenszel Methods for Differential Item Functioning Detection

Ángel M. Fidalgo
Jaqueline M. Madeira
University of Oviedo, Spain

Mantel-Haenszel methods comprise a highly flexible methodology for assessing the degree of association between two categorical variables, whether they are nominal or ordinal, while controlling for other variables. The versatility of Mantel-Haenszel analytical approaches has made them very popular in the assessment of the differential functioning of both dichotomous and polytomous items. Up to now, researchers have limited the use of Mantel-Haenszel statistics to analyzing contingency tables of dimensions 2×2 (by means of the Mantel-Haenszel chi-square statistic) and of dimensions of $2 \times C$ (by means of either the generalized Mantel-Haenszel test or Mantel's test). The main objective of this article is to illustrate a unified framework for the analysis of differential item functioning using the Mantel-Haenszel methods. This is done by means of the generalized Mantel-Haenszel statistic for the analysis of the general case of Q contingency tables with dimensions $R \times C$. Moreover, with the new formulation in consideration, this article reviews the most recent research on differential item functioning and suggests new applications and research lines in relation to the statistics proposed.

Keywords: *Cochran-Mantel-Haenszel methods; generalized Mantel-Haenszel statistics; differential item functioning; polytomous items*

The Mantel-Haenszel methods are one of the most popular nonparametric differential item functioning (DIF) detection procedures. In contrast to parametric procedures, these methods do not require a specific form of item response function or large samples and have been applied for detecting DIF in both dichotomous and polytomous items (see review in Camilli & Shepard, 1994; Clauser & Mazor, 1998; Millsap & Everson, 1993; Penfield & Lam, 2000; Potenza & Dorans, 1995). It was Holland and Thayer (1988) who proposed analyzing DIF in dichotomous

Authors' Note: The writing of this article began during the first author's stay at the Instituto Brasileiro de Pós-Graduaçáo e Extensáo (IBPEX). My sincere thanks to the IBPEX and to Dr. J. R. Facion for the invitation. This research was funded by a grant from the Spanish Ministry of Science and Education (research project SEJ2006-07491). Please address correspondence to Ángel M. Fidalgo, Facultad de Psicología, Plaza de Feijoo, s/n, 33003 Oviedo, Spain; e-mail: fidalgo@uniovi.es.

items using the statistic developed by Mantel and Haenszel (1959) for the analysis of sets of Q contingency tables of dimension 2×2 (denoted by Q: 2×2), where each table corresponds to a level of the covariable or matching variable ($h = 1, \ldots, Q$), the rows correspond to the factor levels ($i = 1, \ldots, R = 2$), and the columns correspond to the levels of the response variable ($j = 1, \ldots, C = 2$).

Abundant research has revealed with great clarity the possibilities and limitations of this methodology for the detection of DIF. For example, a series of theoretical and simulation studies have established the relationship between different dichotomous models of item response theory (IRT) and the Mantel-Haenszel (MH) procedure. Specifically, it is well known that for uniform DIF the MH DIF parameter transformed to the "delta scale" is linearly related to the differences in item difficulty between the reference and focal group only in the one-parameter logistic model (1PLM) and the two-parameter logistic model (2PLM; Camilli & Penfield, 1997; Donoghue, Holland, & Thayer, 1993; Roussos, Schnipke, & Pashley, 1999). Furthermore, it is known that only when all items follow the Rasch model does matching on observed raw scores constitute a sufficient statistic for the latent trait parameter (Fischer, 1995; Holland & Thayer, 1998; Zwick, 1990). Nevertheless, even when the above assumptions are not fulfilled and, for example, the test items follow the three-parameter logistic model (3PLM), the MH procedure has proved to be effective in a wide variety of situations (Allen & Donoghue, 1996; Donoghue et al., 1993; Roussos & Stout, 1996; Shealy & Stout, 1993; Uttaro & Millsap, 1994). Also well established is its high power for detecting uniform DIF and its sensitivity for detecting some types of nonuniform DIF (Hidalgo & López-Pina, 2004; Mazor, Clauser, & Hambleton, 1994; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993), its capacity for detecting DIF with small sample sizes (Camilli & Smith, 1990; Fidalgo, Ferreres, & Muñiz, 2004; Fidalgo, Hashimoto, Bartram, & Muñiz, 2007; Mazor, Clauser, & Hambleton, 1992; Muñiz, Hambleton, & Xing, 2001; Parshall & Miller, 1995), and the advantages of using two-stage or iterative purification procedures for purifying the matching variable (Clauser, Mazor, & Hambleton, 1993; Fidalgo, Mellenbergh, & Muñiz, 2000; Narayanan & Swaminathan, 1994, 1996; Rogers & Swaminathan, 1993; Wang & Su, 2004a). It is also fair to point out that substantial latent trait distribution differences of the reference and focal groups yield a highly inflated Type I error (Clauser et al., 1993; Donoghue et al., 1993; Uttaro & Millsap, 1994; Zwick, 1990), especially when the procedure is applied to items that do not fit the 1PLM (Penny & Johnson, 1999).

In the case of polytomous items, generalizations of the MH chi-square statistic (χ^2_{MH}) have also been used for detecting the DIF, always reduced to the case of Q: $2 \times C$ contingency tables: the generalized Mantel-Haenszel test (GMH; Mantel & Haenszel, 1959; Zwick, Donoghue, & Grima, 1993a) and the Mantel test (Mantel, 1963; Zwick et al., 1993a). As discussed in more detail in the following sections, the GMH treats the response categories as nominal data, whereas the Mantel test considers the ordinal nature of the response categories in polytomous items.

Although there are excellent reviews of the MH methods (Fidalgo, 2005; Kuritz, Landis, & Koch, 1988; Stokes, Davis, & Koch, 2000), to date, none of them have been carried out from a DIF perspective. Thus, this article aims to show a unified framework for the analysis of DIF using the MH methods. In particular, the first section shows the generalized Mantel-Haenszel statistic for the analysis of Q: $R \times C$ contingency tables proposed by Landis, Heyman, and Koch (1978). In the next section, we use this framework for analyzing the results of studies on DIF that have used generalized MH methods. Finally, new applications and future research lines based on this methodology are discussed, providing some calculation examples.

Generalized Mantel-Haenszel Statistics

When factor and response variables are reported on categorical measurement scales, either nominal or ordinal, the resulting data can be summarized in contingency tables, and the methods based on the work of Cochran (1954) and Mantel and Haenszel (1959) are commonly used for their analysis. In a general way, it can be said that these methods provide significance tests and, in some cases, tests of two-way association that control for the effects of covariates. The null hypothesis (H_0) they test is that of no partial association, which establishes that in each one of the strata of the covariates, the response variable is distributed randomly with respect to the levels of the factor.

The most common statistic, the MH chi-square (χ^2_{MH}), is applied to the particular case of sets of contingency tables of dimension 2×2. Fortunately, from the outset, various extensions have been proposed for these statistics (Mantel, 1963; Mantel & Haenszel, 1959), all of them particular cases of the analysis of sets of contingency tables with dimensions Q: $R \times C$. The data structure for this general contingency table is shown in Table 1.

Under the H_0 of no partial association and the assumption that the marginal totals $N_{hi\cdot}$ and $N_{h\cdot j}$ are fixed, the observed frequencies in each table (n_{hij}) follow the multiple hypergeometric probability model:

$$\Pr(n_{hij}|H_0) = \frac{\prod_{i=1}^{R} N_{hi\cdot}! \prod_{j=1}^{C} N_{h\cdot j}!}{N_{h\cdot\cdot}! \prod_{i=1}^{R}\prod_{j=1}^{C} n_{hij}!}. \quad (1)$$

In sets of tables 2×2, the multiple hypergeometric distribution is reduced to the hypergeometric distribution

$$\Pr(n_{hij}|H_0) \frac{N_{h1\cdot}!N_{h2\cdot}!N_{h\cdot 1}!N_{h\cdot 2}!}{N_{h\cdot\cdot}!n_{h11}!n_{h12}!n_{h21}!n_{h22}!},$$

Table 1
Data Structure in the *h*th Stratum

	Response Variable Categories						
Factor Levels	1	2	·	j	·	C	Total
1	n_{h11}	n_{h12}	·	n_{h1j}	·	n_{h1C}	$N_{h1\cdot}$
2	n_{h21}	n_{h22}	·	n_{h2j}	·	n_{h2C}	$N_{h2\cdot}$
⋮	⋮	⋮	⋮	⋮	.	⋮	⋮
i	n_{hi1}	n_{hi2}	·	n_{hij}	·	n_{hiC}	$N_{hi\cdot}$
⋮	⋮	⋮	⋮	⋮	·	⋮	⋮
R	n_{hR1}	n_{hR2}	·	n_{hRj}	·	n_{hRC}	$N_{hR\cdot}$
Total	$N_{h\cdot 1}$	$N_{h\cdot 2}$	·	$N_{h\cdot j}$	·	$N_{h\cdot C}$	$N_{h\cdot\cdot}$

which was used by Mantel and Haenszel (1959) to formulate the famous χ^2_{MH} statistic (see Fidalgo, 2005).

In the general case, the H_0 of no association will be tested against different alternative hypotheses (H_1) that will be a function of the scale on which factor and response are measured. Thus, we shall have a variety of statistics that will serve for detecting the general association (both variables are nominal), mean score differences (factor is nominal and response ordinal), and linear correlation (both variables are ordinal).

The standard generalized Mantel-Haenszel test is defined in terms of matrices by Landis et al. (1978) as

$$Q_{GMH} = \left\{\sum_{h=1}^{Q} (\mathbf{n}_h - \mathbf{m}_h)' \mathbf{A}_h'\right\} \left\{\sum_{h=1}^{Q} \mathbf{A}_h \mathbf{V}_h \mathbf{A}_h'\right\}^{-1} \left\{\sum_{h=1}^{Q} \mathbf{A}_h (\mathbf{n}_h - \mathbf{m}_h)\right\}. \tag{2}$$

where $\mathbf{n}_h$, $\mathbf{m}_h$, $\mathbf{V}_h$, and $\mathbf{A}_h$ are, respectively, the vector of observed frequencies, the vector of expected frequencies, the covariances matrix, and a matrix of linear functions defined in accordance with the H_1 of interest. From Table 1, these vectors and matrices are defined as

$$\mathbf{n}_h = (n_{h11}, n_{h21}, \ldots, n_{hRC})' \qquad (CR \times 1),$$

$$\mathbf{m}_h = N_{h\cdot\cdot}(\mathbf{p}_{h\cdot *} \otimes \mathbf{p}_{h*\cdot}) \qquad (CR \times 1),$$

$$\mathbf{V}_h = N_{h\cdot\cdot}^2/(N_{h\cdot\cdot} - 1)\{(\mathbf{D}_{p_{h\cdot *}} - \mathbf{p}_{h\cdot *}\mathbf{p}_{h\cdot *}') \otimes (\mathbf{D}_{p_{h*\cdot}} - \mathbf{p}_{h*\cdot}\mathbf{p}_{h*\cdot}')\} \qquad (CR \times CR),$$

where $\mathbf{p}_{h*\cdot}$ and $\mathbf{p}_{h\cdot *}$ are, respectively, $(R \times 1)$ and $(C \times 1)$ vectors with the marginal row proportions ($p_{hi\cdot} = N_{hi\cdot}/N_{h\cdot\cdot}$) and the marginal column proportions ($p_{h\cdot j} = N_{h\cdot j}/N_{h\cdot\cdot}$), $\otimes$ denoting the Kronecker product multiplication, $\mathbf{D}_{p_{h\cdot *}}$ is a $(C \times C)$ diagonal matrix with elements of the vector $\mathbf{p}_{h\cdot *}$ on its main diagonal, and $\mathbf{D}_{p_{h*\cdot}}$ is an $(R \times R)$ diagonal matrix with elements of the vector $\mathbf{p}_{h*\cdot}$ on its main diagonal.

As it has been pointed out, depending on the measurement scale of factor and response, expression (2) will be resolved via definition of the matrix $\mathbf{A}_h$ ($\mathbf{A}_h = \mathbf{C}_h \otimes \mathbf{R}_h$) in a different statistic for detecting each H_1. Briefly, these are as follows.

$Q_{GMH(1)}$ or the Generalized Nominal MH statistic

When the variable row and the variable column are nominal, the H_1 specifies that the distribution of the response variable differs in nonspecific patterns across levels of the row factor. Here, $\mathbf{R}_h = [\mathbf{I}_{R-1}, -\mathbf{J}_{R-1}]$ and $\mathbf{C}_h = [\mathbf{I}_{C-1}, -\mathbf{J}_{C-1}]$, where $\mathbf{I}_{R-1}$ is an $(R-1 \times R-1)$ identity matrix, and $\mathbf{J}_{R-1}$ is an $(R-1 \times 1)$ vector of ones. Thus, the dimension of $\mathbf{R}_h$ will be $(R-1 \times \mathrm{R})$. Similarly, $\mathbf{I}_{C-1}$ is a $(C-1 \times C-1)$ identity matrix, and $\mathbf{J}_{C-1}$ is a $(C-1 \times 1)$ vector of ones. (An example of the construction of this and successive linear operator matrices can be found in Fidalgo, 2005). Under H_0, $Q_{GMH(1)}$ follows approximately a chi-square distribution with degrees of freedom $(df) = (R-1)(C-1)$. When $R = C = 2$, $Q_{GMH(1)}$ is identical to the statistic proposed by Birch (1964) and identical to the χ^2_{MH} statistic, except for the lack of the continuity correction. For the special case of 2 factor levels, $Q_{GMH(1)}$ is identical to the generalized Mantel-Haenszel test (GMH) proposed by Mantel and Haenszel (1959).

$Q_{GMH(2)}$ or the Generalized Ordinal MH statistic

When only the variable column is ordinal, the H_1 establishes that the mean responses differ across the factor levels, $\mathbf{R}_h$ being the same as that used in the previous case and $\mathbf{C}_h = (c_{h1}, \ldots, c_{hC})$ being a $(1 \times C)$ vector, where c_{hj} is an appropriate score reflecting the ordinal nature of the jth category of response for the hth stratum. Selection of the values of $\mathbf{C}_h$ admits different possibilities that are well described in Koch, Imrey, Singer, Atkinson, and Stokes (1985) and Landis et al. (1978). Under H_0, $Q_{GMH(2)}$ has approximately a chi-square distribution with $df = (R-1)$. For the special case of 2 factor levels, $Q_{GMH(2)}$ is identical to the extended MH test proposed by Mantel (1963). Moreover, it should be noted that the square root of the Mantel test is equal to Cox's (1958) statistic for testing the null hypothesis that the noniterative estimator of the noncentrality parameter (β) of the multivariate hypergeometric distribution is equal to 0.

As it has been pointed out, calculation of the $Q_{GMH(2)}$ statistic requires selecting the scores ($\mathbf{C}_h$) that will be applied to the response variable to compute the row mean scores $\left[\bar{y}_{hi} = \sum_{j=1}^{C} (c_{hj} n_{hij} / N_{hi\cdot})\right]$ used for comparing the factors across strata. Options include integer scores, rank scores, ridit scores, modified ridit scores, and log-rank scores. In this situation, what scores should be selected for the calculation of $Q_{GMH(2)}$? When the choice of scores is not obvious, Graubard and

Korn (1987) recommend choosing integer (or equally spaced) scores, and Agresti (1990, p. 294) advocates analyzing the data with a few reasonable sets of scores to see if the choice might affect the results.

$Q_{GMH(3)}$ or the Generalized Correlation MH statistic

If both variables are ordinal, the H_1 establishes the existence of a linear progression in the mean responses across the levels of the factor. In this case, $\mathbf{C}_h$ can be defined as the same as that for the mean responses difference and $\mathbf{R}_h = (r_{h1}, \ldots, r_{hR})$ is a $(1 \times R)$ vector, where r_{hi} is an appropriate score reflecting the ordinal nature of the ith factor level for the hth stratum. Under H_0, $Q_{GMH(3)}$ has approximately a chi-square distribution with $df = 1$.

This statistic is identical to the correlation statistic proposed by Mantel (1963) and Birch (1965). When marginal rank or ridit scores are assigned to both columns ($\mathbf{C}_h$) and rows ($\mathbf{R}_h$) of each table, with midranks assigned for ties, $Q_{GMH(3)}$ is equivalent to the Spearman rank correlation, conditioning on the levels of the strata. More specifically, when there is only one stratum, $Q_{GMH(3)}$ is equal to $(n-1)r_S^2$, with $n = \sum_{h=1}^{Q} N_{h..}$ and r_S being the Spearman rank correlation coefficient. If in the same situation we were to use integer scores, $Q_{GMH(3)}$ would be equal to $(n-1)r^2$, r being the Pearson correlation coefficient.

It should be noted how the successive H_1s specify more and more restrictive patterns of association, so that each statistic increases the statistical power with respect to the previous ones for detecting its particular pattern of association. For example, $Q_{GMH(1)}$ can detect linear patterns of association, but it will do so with less power than $Q_{GMH(3)}$. Furthermore, the increase in power of $Q_{GMH(3)}$ compared to $Q_{GMH(1)}$ is achieved at the cost of an inability to detect more complex patterns of association. Obviously, when $C = R = 2$, $Q_{GMH(1)} = Q_{GMH(2)} = Q_{GMH(3)} = \chi^2_{MH}$, except for the lack of the continuity correction.

The statistics seen up to now are all inferential tests. In the case of rejecting the H_0, the following step is to determine the degree of association between factor and response. Although MH methods have satisfactorily resolved the testing of the H_0 in the general case, to date there is no estimator of the degree of association that is completely generalizable for Q: $R \times C$ tables. For the particular case of Q: 2×2 contingency tables, Mantel and Haenszel (1959) proposed the well-known common odds ratio estimator ($\hat{\alpha}_{MH}$). But there are generalizations, always complex, of $\hat{\alpha}_{MH}$ to Q: $2 \times C$ tables (Greenland, 1989; Liu & Agresti, 1996; Mickey & Elashoff, 1985; Yanagawa & Fujii, 1995). The Greenland and Yanagawa-Fujii (a correction of the Greenland common odds ratio) estimator yields $C-1$ estimators of the common odds ratio, whereas the Liu-Agresti provides a single estimator of the common odds ratio ($\hat{\psi}_{LA}$). The latter is a generalization of $\hat{\alpha}_{MH}$, so that in the particular case of $C = 2$, $\hat{\psi}_{LA} = 1/\hat{\alpha}_{MH}$. Liu and Agresti (1996) also provide an estimation of the variance of $\log(\hat{\psi}_{LA})$ that is a generalization of the variance estimator of $\hat{\alpha}_{MH}$

developed by Robins, Breslow, and Greenland (1986). For a complete exposition of the Liu-Agresti statistic for DIF detection, the reader is recommended to consult Penfield and Algina (2003).

Generalized Mantel-Haenszel Methods in the DIF Literature

Although the first studies on DIF in polytomous items only appeared in the early 1990s, there has already been sufficient research for a number of authors to undertake theoretical reviews (Penfield & Lam, 2000; Potenza & Dorans, 1995). As we know, polytomous items are those that present more than two response categories; a prototypical case would be Likert-type items. The analysis of DIF in this type of item by means of MH methods has been confined to two particular cases of the statistics presented in the previous section: the generalized Mantel-Haenszel test (Mantel & Haenszel, 1959) and the Mantel test (Mantel, 1963).

Table 2 summarizes the main characteristics of the studies on DIF that have used some type of generalized MH statistic. As can be seen, the majority of the studies were designed to replicate educational tests that contain both multiple-choice items (using dichotomous IRT models to generate data) and essay items (using polytomous IRT models to generate data). Only the works by Wang and Su (2004b) and by Kristjansson, Aylesworth, McDowell, and Zumbo (2005) are designed to replicate tests with only polytomous items, such as constructed-response test or scales that use Likert-type items. This variable that in principle could limit the generalization of results to the type of test simulated, appears to have little influence on the behavior of the MH statistics tested. Thus, for example, it has been shown that the Mantel test yields higher power than GMH under constant DIF patterns (the magnitude of DIF is constant across score categories), regardless of whether the majority of the test items are dichotomous (Chang, Mazzeo, & Roussos, 1996; Zwick et al., 1993a; Su & Wang, 2005) or all are polytomous (Wang & Su, 2004b).

Inevitably, there are a series of parallels between the MH procedure for dichotomous items and the generalized MH statistics. The first of these is the advantage—as is also the case for the MH procedure—of including the item under study in the matching score (Zwick et al., 1993a). Second, these techniques perform best when the matching test items follow the family of Rasch models, such as the partial credit model. When this is not the case and, for instance, the items follow the graded response model (which is an extension of the 2PLM) and the mean latent trait differences between groups is large, these methods show inflated Type I error rates (Su & Wang, 2005; Wang & Su, 2004b). The same occurs when the discrimination parameter of the studied item differs from the average discrimination parameters of the test items (Chang et al., 1996). Third, Mantel's test, like χ^2_{MH} in the case of dichotomous items, has low power for detecting nonuniform DIF, but this

is not the case with GMH (Kristjansson et al., 2005). Finally, in a comprehensive study, Wang and Su (2004b) showed that two-stage and iterative purification procedures do not improve the efficiency of the generalized MH statistics as they do for the χ^2_{MH} statistic. They found that under the pattern of balanced DIF (the magnitude of DIF is balanced across score categories) within the item, or constant within the item but balanced within the test, there is no substantial improvement, though improvement does occur in other conditions (Su & Wang, 2005).

In all the cases of application of generalized MH statistics, the group was the factor variable and the response variable was the polytomous item response, except in Penfield (2001). In that study, the GMH statistic was used for detecting DIF among multiple groups, the dichotomous item response being the factor variable and the group variable the response variable ($C \geq 2$, one reference group and from 1 to 4 focal groups). That is, generalized MH statistics have been used in studies on DIF either to detect the DIF in dichotomous items across several groups simultaneously or to determine whether there is differential functioning of the item between two groups (focal and reference) across the different categories of a polytomous item. In the following section, we discuss some new application possibilities for generalized MH statistics.

New Applications and Future Research Lines

To date, the generalized MH statistics used in research on DIF have confined themselves to the analysis of Q contingency tables of dimensions $2 \times C$, and always by means of the statistics $Q_{GMH(1)}$ and $Q_{GMH(2)}$. The first and natural extension of this methodology will be therefore the analysis of DIF in multiple groups ($R > 2$), that is, the application of the statistics presented to the general case of tables of dimensions $R \times C$. In the case of dichotomous items, the simplest way of assessing the DIF using $Q_{GMH(1)}$ is to use the groups as factor and item response as response variable. That is, tables of dimensions $R \times 2$ are to be analyzed. It should be pointed out that the analytical strategy proposed by Penfield (2001), which takes as factor the item response and as response variable the groups (analyzing tables of dimensions $2 \times C$), will provide the same results. Penfield's proposal, though effective, is sufficiently counterintuitive for us to assume that the use of the GMH statistic in this situation is determined solely by its customary use in the literature on DIF.

As it has been stated, the GMH statistic is a particular case of $Q_{GMH(1)}$, so that it would be more advisable to apply the latter in the form initially described. In the case of polytomous items, depending on whether the response variable (the items) is nominal or ordinal, one shall use the statistics $Q_{GMH(1)}$ or $Q_{GMH(2)}$, respectively. However, it should be noted that irrespective of the measurement scale of the item, $Q_{GMH(1)}$ is the best option for detecting balanced DIF patterns and $Q_{GMH(2)}$ is the best option for detecting constant DIF patterns (Fidalgo & Bartram, 2008). Now, when should

Table 2
Differential Item Functioning (DIF) Studies Using Generalized Mantel-Haenszel (MH) Methods

Studies[a]	Inferential Procedures[b]	DIF Effect Size Estimator	DIF Patterns[c]	Generation Model[d]
Welch & Hoover (1993)	$Q_{GMH(2)}$ vs. HW1 and HW3		Constant	3PLM/partial credit model
Zwick, Donoghue, & Grima (1993a)	$Q_{GMH(1)}$ and $Q_{GMH(2)}$	SMD	Constant, balanced and unbalanced	3PLM/partial credit model
Zwick & Thayer (1996)	$Q_{GMH(2)}$	SMD	Constant	3PLM/partial credit model
Chang, Mazzeo, & Roussos (1996)	$Q_{GMH(2)}$ vs. SMD[e] and POLY-SIBTEST		Constant, balanced, and unbalanced	3PLM/partial credit model and generalized partial credit model
Zwick, Thayer, & Mazzeo (1997)	$Q_{GMH(2)}$ vs. SMD and SIBTEST	SMD and SIBTEST DIF estimator	Constant	3PLM/generalized partial credit model
Ankenmann, Witt, & Dunbar (1999)	$Q_{GMH(2)}$ vs. likelihood ratio goodness-of-fit statistic		Constant and balanced	3PLM/graded response model
Camilli & Congdon (1999)	$Q_{GMH(2)}$[f] vs. logistic regression	Cox and logistic regression coefficients	No DIF	3PLM/generalized partial credit model
Penfield (2001)	$Q_{GMH(1)}$ and χ^2_{MH}		Constant	3PLM
Penfield & Algina (2003)	$Q_{GMH(2)}$[f] vs. Liu-Agresti estimator	Liu-Agresti estimator	Constant and unbalanced	3PLM/generalized partial credit model and graded response model
Meyer, Huynh, & Seaman (2004)	$Q_{GMH(2)}$ vs. exact tests[g]	SMD		
Wang & Su (2004b)	$Q_{GMH(1)}$ and $Q_{GMH(2)}$		Constant and balanced	Partial credit model and graded response model

(continued)

Table 2 (continued)

Studies[a]	Inferential Procedures[b]	DIF Effect Size Estimator	DIF Patterns[c]	Generation Model[d]
Su & Wang (2005)	$Q_{GMH(1)}$ and $Q_{GMH(2)}$ vs. logistic discriminant function analysis		Constant, balanced, and unbalanced	3PLM/partial credit model and graded response model
Kristjansson, Aylesworth, McDowell, & Zumbo (2005)	$Q_{GMH(1)}$ and $Q_{GMH(2)}$ vs. logistic discriminant function analysis and unconstrained cumulative logits ordinal logistic regression		Constant and nonuniform	Generalized partial credit model

a. All are simulation studies except that of Meyer et al. (2004).
b. Generalized MH statistics were always used for the case of Q: $2 \times C$ tables, that is, the GMH statistics and the Mantel test.
c. In all the studies, except in Kristjansson et al. (2005), a uniform DIF was generated.
d. In the cases in which there are dichotomous and polytomous models, the tests were generated with both types of items.
e. SMD = standardized mean difference.
f. The authors used a statistic equivalent to Mantel test, Cox's β noncentrality parameter test.
g. The exact tests were the Wilcoxon Rank Sum tests and the Waerden Normal Scores test.

$Q_{GMH(3)}$ be used? As pointed out, the $Q_{GMH(3)}$ statistic tests the hypothesis of a linear relationship between the factor (the groups) and the response variable. Thus, in all those cases in which this hypothesis is viable, $Q_{GMH(3)}$ should be used, because with 1 degree of freedom specifically focused on the alternative of a monotone relationship between group and item response, being the most powerful statistic in this situation. For example, it is possible to consider the use of $Q_{GMH(3)}$ for examining hypotheses on the causes of DIF within a confirmatory framework (Roussos & Stout, 1996; Ryan & Chiu, 2001; Walker & Beretvas, 2001). Walker and Beretvas (2001), in showing that open-ended items measuring mathematical communication function differentially in favor of examinees who are proficient writers, carried out confirmatory DIF analysis using poly-SIBTEST. In this situation we could also use the $Q_{GMH(3)}$ statistic to check this hypothesis. To this end, we would form several groups according to examinees' writing proficiency (the spurious variable) and analyze the relationship between the response to suspicious items (which are assumed to measure the spurious and principal variables) and the groups by means of $Q_{GMH(3)}$. It should be pointed out that Walker and Beretvas (2001) did not consider writing proficiency as a spurious variable but rather as a valid dimension.

A research line as yet unexplored—partly because since the formulation of the statistics this matter has been ignored—is the effect of the choice of scores assigned to the ordinal variables on the $Q_{GMH(2)}$ and $Q_{GMH(3)}$ statistics. As Cochran (1954) and Agresti (1990) point out, if the set scores do not adequately reflect the numerical scale that underlies the ordinal variable, the test will not be sensitive. Studies on DIF have routinely used integer scores (Ankenmann, Witt, & Dunbar, 1999; Wang & Su, 2004b; Zwick & Thayer, 1996), with only Meyer, Huynh, and Seaman (2004) following Mantel's (1963) example of using ridit scores. Such uniformity together with the nonexplicit formulation of the selection may lead to the illusion that the statistics are not sensitive to the choice of scores, when in fact nothing could be further from the truth. For example, as early as 1987, Graubard and Korn argued that rank scores can be a poor choice when the column marginal is far from uniformly distributed. On the other hand, log-rank scores can be useful when the distribution is L-shaped and there is greater interest in factor differences for response levels with higher values (Koch, Imrey, et al., 1985; Koch, Sen, & Amara, 1985). This is a research line of undoubted interest.

Another area of research that should be developed is that relative to the measures of DIF effect size. Of the estimators associated with the MH methods, the proposals for measures of DIF are the Yanagawa-Fujii (Zwick, Donoghue, & Grima 1993b) and the Liu-Agresti estimators (Penfield & Algina, 2003), the latter presenting the greatest advantages. However, up to now, there have been insufficient comparative studies between these estimators and others, such as standardized mean difference or the Cox and Logistic regression coefficients (Camilli & Congdon, 1999). Furthermore, no classificatory schemes have been developed that combine statistical significance with practical significance using these statistics, as in the case of dichotomous

items (Zieky, 1993). Decision making in relation to DIF requires both powerful and robust inferential tests and reliable measures of DIF magnitude.

Example

Below we illustrate the calculation of Mantel-Haenszel statistics by means of data simulated to provide an example of one of the research areas previously proposed. With this purpose we carried out a brief simulation study. An artificial test was constructed with 20 dichotomous items and 4 polytomous items with four ordinal response categories. The item parameters, typical of those found in applied testing settings, are identical to those used by Penfield and Algina (2003). The specified polytomous item parameters of the reference group are as follows: Item 1 ($a=0.8$, $b_{i1}=-1.0$, $b_{i2}=0.0$, $b_{i3}=1.0$); Item 2 ($a=0.8$, $b_{i1}=-1.2$, $b_{i2}=-0.5$, $b_{i3}=1.0$); Item 3 ($a=1.3$, $b_{i1}=-1.0$, $b_{i2}=0.0$, $b_{i3}=1.0$); Item 4 ($a=1.3$, $b_{i1}=-1.2$, $b_{i2}=-0.5$, $b_{i3}=1.0$). The dichotomous item parameters can be found in Chang et al. (1996).

The 20 dichotomous items were generated from a three-parameter logistic IRT model (3PLM). In the 3PLM, the probability of a correct response for an examinee with latent trait θ is defined by

$$P_i(\theta)=c_i+(1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}},$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is a pseudo guessing parameter, and D is a scaling factor equal to 1.7. The 4 polytomous items were generated using the graded response model (GRM). In the GRM, the probability of scoring x and above on item i with K categories (from $k=1$ to K), given θ, is defined by

$$P^*_{i(x\geq k)}(\theta)=\frac{e^{Da_i(\theta-b_{ik-1})}}{1+e^{Da_i(\theta-b_{ik-1})}},$$

where a_i is the item discrimination parameter, b_{ik} is the kth item difficulty parameter, and D is the scaling factor. By definition, $P^*_{i(x\geq 1)}(\theta)=1$ and $P^*_{i(x\geq K+1)}(\theta)=0$, so the probability of scoring in the category k is equal to $P_{ik}=P^*_{i(x\geq k)}(\theta)-P^*_{i(x\geq k+1)}(\theta)$.

To generate the DIF, the b_{ik} parameters of the 4-point items were changed for the focal group according to the following equations:

$b_{ikF}=b_{ikR}+s$, $k=1,2,3$, and $s=0.20$ (constant DIF pattern),
$b_{ikF}=b_{ikR}+s$, $k=3$, and $s=0.40$ (shift-high DIF pattern),

where s is equal to the magnitude of DIF manipulated. In the No DIF condition, the b_{ik} parameters were the same in both groups.

For each of the three DIF patterns (No DIF, constant DIF, shift-high DIF), 1,000 data sets were generated using the GAUSS program (V.3.1.4). The ability distribution used to generate the response data set were the same in both the reference (500 examinees) and the focal group (500 examinees): $N(0, 1)$. In short, 3,000 different tests were analyzed with the $Q_{GMH(1)}$ and $Q_{GMH(2)}$ statistics (using integer and log-rank scores) using a significance level of .05. Log-rank scores are computed from the general contingency table (shown in Table 1) using

$$c_{hj} = 1 - \sum_{k=1}^{j} \left(\frac{N_{h \cdot k}}{\sum_{m=k}^{C} N_{h \cdot m}} \right). \tag{3}$$

With the aim of comparing the different statistics in an optimum situation, we used for calculation of the matching variable (total test score) the 20 dichotomous non-DIF items plus the polytomous item under analysis. To calculate the MH statistics a computer program was written in the GAUSS programming language. This software is available on request from the authors.

To illustrate the calculation of the Mantel-Haenszel statistics we use the responses in one of the simulated tests to Item 4 when it presents a shift-high DIF pattern. As is customary in DIF studies, we used as matching variable the total score on the test. The factor is group (focal or reference) and the response variable is item category. Table 3 shows the frequencies observed, corresponding to each of the seven 2×4 contingency tables that we shall use for calculating the generalized MH statistics. Table 4 shows the results obtained on calculating the MH statistics together with the linear operator matrix defined in accordance with the H_1s of general association ($Q_{GMH(1)}$) and mean row scores difference ($Q_{GMH(2)}$). As expected, it is clear that the use of log-rank scores provides a more sensitive result ($Q_{GMH(2)-Log\text{-}rank} = 4.613$; $df = 1$; $p = .032$) than the use of integer scores ($Q_{GMH(2)-Integer} = 2.088$; $df = 1$; $p = .148$), on giving greater weight to the response category of the item in which DIF is present.

The test power obtained in the simulation study described above corroborates these results (see Table 5). As already established, $Q_{GMH(2)}$ increased the statistical power with respect to $Q_{GMH(1)}$ for testing whether the mean responses differ across the factor levels. Consequently, in the case that the response variable is ordinal and the pattern of association fits the hypothesis described above, as occurs when DIF is constant, $Q_{GMH(2)}$ should be the statistic of choice (see Table 5). On the other hand, the shift-high DIF pattern condition serves to illustrate how $Q_{GMH(1)}$ offers the possibility of detecting more complex patterns of association than $Q_{GMH(2)}$. Furthermore, in such a situation, it is evident that the choice of scores on calculating $Q_{GMH(2)}$ has a clear effect on its power.

Table 3
Frequency Tables With the Responses Given by Focal and Reference Groups to a 4-Point Item With Differential Item Functioning in the Highest Category

		Response Item Categories			
Ability Level	Group	1	2	3	4
$h = 1$					
	Reference	9	0	0	0
	Focal	12	0	0	0
	Total	21	0	0	0
	Log-rank	0.511	0.178	–0.322	–1.322
$h = 2$					
	Reference	30	12	2	0
	Focal	31	7	3	0
	Total	61	19	5	0
	Log-rank	0.642	0.249	–0.251	–1.251
$h = 3$					
	Reference	38	35	25	0
	Focal	31	35	24	0
	Total	69	70	49	0
	Log-rank	0.816	0.522	0.022	–0.978
$h = 4$					
	Reference	11	31	62	1
	Focal	14	45	67	3
	Total	25	76	129	4
	Log-rank	0.893	0.530	–0.440	–1.440
$h = 5$					
	Reference	3	18	90	27
	Focal	4	12	93	13
	Total	7	30	183	40
	Log-rank	0.973	0.854	0.034	–0.966
$h = 6$					
	Reference	0	2	27	47
	Focal	0	0	47	30
	Total	0	2	74	77
	Log-rank	1.000	0.987	0.497	–0.503
$h = 7$					
	Reference	0	0	2	28
	Focal	0	0	3	26
	Total	0	0	5	54
	Log-rank	1.000	1.000	0.915	–0.085

Note: An interval of three units in the scale score was used for matching examines. Log-rank scores are computed using Equation 3.

Table 4
Results of the Mantel-Haenszel Statistics With the Respective Linear Operator Matrices

Statistic	Score Type	Value	*df*	*p*	$A_h = C_h \otimes R_h$
$Q_{GMH(1)}$		10.178	3	.017	$A_h^{a} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \otimes [1 \quad -1]$
$Q_{GMH(2)}$	Integer	2.088	1	.148	$A_h = [1 \quad 2 \quad 3 \quad 4] \otimes [1 \quad -1]$
	Log-rank	4.613	1	.032	$A_1^{b} = [0.511 \quad 0.178 \quad -0.322 \quad -1.322] \otimes [1 \quad -1]$
					$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$
					$A_7 = [1.000 \quad 1.000 \quad 0.915 \quad -0.085] \otimes [1 \quad -1]$
$Q_{GMH(3)}{}^{c} = Q_{GMH(2)}$					

a. For the calculation of both $Q_{GMH(1)}$ and $Q_{GMH(2)}$ when integer scores are used, the linear operator matrices are the same in all stratum levels.
b. The log-rank scores used in C_h were shown in Table 3.
c. When, as in this case, $R = 2$, $Q_{GMH(3)} = Q_{GMH(2)}$.

Table 5
Type I Error Rate and Statistical Power

		Generalized Mantel-Haenszel Statistics		
DIF Pattern	Item	$Q_{GMH(1)}$	$Q_{GMH(2)}$-Integer	$Q_{GMH(2)}$-Log-rank
No DIF	1	.047	.043	.042
	2	.055	.054	.052
	3	.045	.040	.038
	4	.049	.047	.042
	Average	.049	.046	.044
Shift-high	1	.821	.213	.371
	2	.758	.224	.399
	3	.936	.338	.508
	4	.917	.369	.573
	Average	.858	.286	.463
Constant	1	.334	.518	.487
	2	.353	.516	.498
	3	.700	.838	.821
	4	.685	.814	.811
	Average	.518	.672	.654

Conclusion

The main objective of this article was to provide a framework for integrating the different MH statistics used in research on DIF, which have shown themselves to be particular cases of the generalized MH test used in the analysis of Q contingency tables with dimensions $R \times C$. This overall perspective allows us to analyze and connect statistics that are apparently different. It also increases the application possibilities of an extremely fertile and robust methodology for DIF analysis, exploring subtle and not-so-subtle differences and providing the ideal framework for developing the research lines discussed above.

References

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

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement*, *33*, 231-251.

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*(4), 277-300.

Birch, M. W. (1964). The detection of partial association, I: The 2 × 2 case. *Journal of the Royal Statistical Society, Series B*, *26*, 313-324.

Birch, M. W. (1965). The detection of partial association, II: The general case. *Journal of the Royal Statistical Society, Series B*, *27*, 111-124.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioural Statistics*, *4*, 323-341.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement*, *34*, 123-139.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Camilli, G., & Smith, J. K. (1990). Comparison of the Mantel-Haenszel test with a randomized and a jackknife test for detecting biased items. *Journal of Educational Statistics*, *15*, 53-67.

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

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31-44.

Clauser, B., Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *6*, 269-279.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 test. *Biometrics*, *10*, 417-451.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society, Series B*, *20*, 215-232.

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

Fidalgo, A. M. (2005). Mantel-Haenszel methods. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1120-1126). Chichester, UK: Wiley.

Fidalgo, A. M., & Bartram, D. (2008). *DIF detection in polytomous items using the generalized ordinal Mantel-Haenszel statistic. What type of scores should be selected?* Manuscript submitted for publication.

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning with small samples. *Educational and Psychological Measurement*, *64*, 925-936.

Fidalgo, A. M., Hashimoto, K., Bartram, D., & Muñiz, J. (2007). Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. *Journal of Experimental Education*, *75*, 293-314.

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research*, *5*, 43-53.

Fischer, G. (1995). Some neglected problems in IRT. *Psychometrika*, *60*, 459-487.

Graubard, B. I., & Korn, E. L. (1987). Choice of column scores for testing independence in ordered 2 × K contingency tables. *Biometrics*, *43*, 471-476.

Greenland, S. (1989). Generalized Mantel-Haenszel estimators for K 2 × J tables. *Biometrics*, *45*, 183-191.

Hidalgo, M. D., & López-Pina J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*, 903-915.

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

Koch, G. G., Imrey, P. B., Singer, J. M., Atkinson, S. S., & Stokes, M. E. (1985). *Analysis of categorial data*. Montreal, Canada: Presses de l'Université de Montreal.

Koch, G. G., Sen, P. K., & Amara, I. A. (1985). Log-rank scores, statistics and test. In N. L. Johnson & S. Kotz (Eds.), *Encyclopedia of statistical sciences* (Vol. 5, pp. 136-142). New York: John Wiley.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*, 935-953.

Kuritz, S. J., Landis, J. R., & Koch, G. G. (1988). A general overview of Mantel-Haenszel methods: Applications and recent developments. *Annual Review of Public Health*, *9*, 123-160.

Landis, J. R., Heyman, E. R., & Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. *International Statistical Review*, *46*, 237-254.

Liu, I.-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics*, *52*, 1223-1234.

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.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement*, *52*, 443-452.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement*, *54*, 284-291.

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004). Exact small-sample differential item functioning for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement*, *41*, 331-344.

Mickey, R. M., & Elashoff, R. M. (1985). A generalization of the Mantel-Haenszel estimator of partial association for 2 × J × K tables. *Biometrics*, *41*, 623-635.

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

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, *1*, 115-135.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*, 315-328.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*, 257-274.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement*, *32*, 302-316.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, *14*, 235-259.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement*, *40*(4), 353-370.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, *19*, 5-15.

Penny, J., & Johnson, R. L. (1999). How group differences in matching criterion distribution and IRT item difficulty can influence the magnitude of the Mantel-Haenszel chi-square DIF index. *Journal of Experimental Education*, *67*, 343-366.

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.

Robins, J., Breslow, N., & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, *42*, 311-324.

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.

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics*, *24*(3), 293-322.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33*, 215-230.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education*, *14*, 73-90.

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

Stokes, M. E., Davis, C. S., & Koch, G. G. (2000). *Categorical data analysis using the SAS system* (2nd ed.). Cary, NC: SAS Institute.

Su, Y.-H., & Wang, W.-C. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning in polytomous items. *Applied Measurement in Education*, *18*, 313-350.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement*, *18*, 15-25.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, *38*, 147-163.

Wang, W.-C., & Su, Y.-H. (2004a). Effect of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*, 113-144.

Wang, W.-C., & Su, Y.-H. (2004b). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, *28*, 450-480.

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

Yanagawa, T., & Fujii, Y. (1995). Projection-method Mantel-Haenszel estimator for K $2 \times$ J tables. *Journal of the American Statistical Association*, *90*, 649-656.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In W.P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum.

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. (1993a). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233-251.

Zwick, R., Donoghue, J. R., & Grima, A. (1993b). *Assessing differential item functioning in performance tests* (ETS Research Rep. No. 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 Behavioural Statistics*, *21*, 187-201.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education*, *10*, 321-344.