APPLIED MEASUREMENT IN EDUCATION, *14*(3), 235–259

Assessing Differential Item Functioning Among Multiple Groups: A Comparison of Three Mantel–Haenszel Procedures

Randall D. Penfield
University of Florida

It is often the case in performing a differential item functioning (DIF) analysis that comparisons are made between a single reference group and multiple focal groups. Conducting a separate test of DIF for each focal group has several undesirable qualities: (a) the Type I error rate will exceed the intended nominal level if the level of significance for each individual test is not appropriately adjusted, (b) the power may not be as high as a single test that assesses DIF among all groups simultaneously, and (c) substantial time and computing resources are required. These drawbacks are potentially avoided by using a procedure that has the capacity to assess DIF across all groups simultaneously. In this study I compare the performance of three methods of assessing DIF across multiple demographic groups; the Mantel–Haenszel chi-square statistic with no adjustment to the alpha level, the Mantel–Haenszel chi-square statistic with a Bonferroni adjusted alpha level, and the Generalized Mantel–Haenszel statistic (GMH) that offers a single test of significance across all groups. Simulations were conducted in which there was a single reference group and 1, 2, 3, and 4 focal groups, having from 1 to all of the focal groups in a given condition experiencing DIF. Additional conditions that were varied included group size, focal group ability distribution, and magnitude of matching criterion contamination. The results suggest that GMH is in general the most appropriate procedure because its Type I error rate remained at the nominal level of 0.05, and its power was consistently among the highest.

Concern for equality in testing during the 1960s and 1970s led to a surge in the development of statistical methodology for item bias detection (Camilli & Shepard,

Requests for reprints should be sent to Randall D. Penfield, Department of Educational Psychology, College of Education, University of Florida, P.O. Box 117047, Gainesville, FL 32611–7047.

1994; Cole, 1993). Incipient statistical investigations into item bias were predicated on the identification of items that displayed unusually large differences in the mean performance between demographic groups relative to that observed for the other items on the test (Angoff, 1972; Cleary & Hilton, 1968). Although these methods enabled the identification of items that were differentially difficult for one group of examinees, they lacked a rigorous method for determining whether differences in group performance on an item were caused by some form of unfairness in the item, or simply differing levels of proficiency in the groups being compared. Modern investigations into item bias control for the confounding effects of differing levels of group proficiency by using the framework of differential item functioning (DIF), which is defined as existing when examinees from different demographic groups perform differently on an item after conditioning on the ability intended to be measured by the test (Dorans & Holland, 1993). The presence of DIF may indicate the existence of a systematic invalidity of the item, placing one group at a disadvantage.

Over the past 2 decades, numerous DIF detection procedures have been developed for both dichotomous and polytomous items (see Camilli & Shepard, 1994; Clauser & Mazor, 1998; Millsap & Everson, 1993; Penfield & Lam, 2000; Potenza & Dorans, 1995). All of these approaches were developed exclusively for the two-group case in which comparisons are made between a base (reference) group and a second (focal) group. It is frequently desirable, however, to assess item bias for several focal groups. Numerous focal groups have been identified as important candidates for DIF investigation: Asian Americans, Blacks, Hispanics, Native Americans, women, and examinees with disabilities (Zieky, 1993). Linn (1993) suggested a further refinement of focal group categories to distinguish among Puerto Ricans, Mexican Americans, Cubans, and other Hispanic groups. The practical need for considering multiple focal groups is highlighted by the presence of numerous studies in the literature examining DIF among multiple ethnic groups (Schmitt, 1988; Schmitt & Dorans, 1990; Zwick & Ercikan, 1989) and multiple languages of administration (Angoff & Sharon, 1974; Ellis & Kimmel, 1992).

Given the prevalence of multiple-group DIF assessments, investigations into item bias would benefit from the availability of statistical procedures that test for DIF simultaneously across multiple groups. Such procedures have three possible advantages over the traditional two-group methods: (a) the power of detecting DIF across multiple groups simultaneously may be greater than that observed in individual pairwise tests, (b) the inflated Type I error rate expected when DIF is tested between multiple pairs of groups is avoided with a single procedure that tests for DIF across all groups simultaneously, and (c) a single test of DIF across all groups provides a more efficient method of assessing DIF than testing each group individually.

AVAILABLE PROCEDURES FOR TESTING MULTIPLE-GROUP DIF

Little research has been devoted to the development methods that can be used to simultaneously assess DIF across multiple groups. Kim, Cohen, and Park (1995) presented a method of assessing DIF across multiple groups that is based on Lord's (1977, 1980) chi-square method for comparing vectors of item response theory (IRT) item parameters between two groups. The statistic developed by Kim et al. (1995), called the Q_j statistic, compares the vectors of item parameters for three or more groups. If, for a given item, the vectors of its parameters differ significantly between groups, then the item characteristic functions will differ across groups, and the item has been shown to function differentially for the groups tested.

The Q_j statistic has several advantages over two-group methods: it permits a single test of significance that may be more powerful than individual tests for each pair of groups compared, and it avoids the increase in Type I error associated with an individual test for each focal group. However, effective application of the Q_j statistic to applied DIF analyses is limited by several factors. First, because it has been shown that sample sizes of approximately 500 are required for stable item parameter estimation for the two-parameter logistic regression IRT model (Hulin, Lissak, & Drasgow, 1982), adequate performance of the Q_j statistic is likely dependent on having moderate to large group sizes. This poses a problem for the Q_j statistic because minority groups often have relatively small sample sizes. Second, IRT parameter estimation procedures are computationally demanding, making it difficult to compute the Q_j statistic for the large number of items commonly obtained during pilot, field, and operational testing. Third, the Q_j statistic does not consider the density of examinees in the sample along the ability continuum, and thus may signal DIF in regions of the ability scale with sparse data. This constraint is known to adversely affect the performance of Lord's chi-square method (Camilli & Shepard, 1994), and likely has a similar implication for the performance of the Q_j statistic. These limitations suggest the need for an alternative DIF detection procedure that simultaneously assesses DIF across multiple focal groups.

THREE MANTEL–HAENSZEL PROCEDURES

One of the most popular procedures for assessing DIF in dichotomous items is the Mantel–Haenszel (MH) procedure, first developed for use in epidemiological research (Mantel & Haenszel, 1959), and later applied to the detection of DIF by Holland and Thayer (1988). Applying the MH procedure to DIF detection begins by grouping examinees according to an estimate of ability (generally the total test

score), and then forming a two-by-two contingency table crossing group membership (reference and focal) and item performance (correct and incorrect) for each level of ability. Let us denote a particular level of ability by k, where $k = 1, 2, \ldots, m$. Then, the MH chi-square statistic can be used to assess the association between group membership and item performance across all m levels of the estimated ability using

$$\mathrm{MH}\chi^2 = \frac{\left| \sum_{k=1}^{m} [A_k - E(A_k)] - \frac{1}{2} \right|^2}{\sum_{k=1}^{m} \mathrm{VAR}(A_k)} \tag{1}$$

where A_k equals the number of correct reference group responses at ability level k,

$$\mathrm{VAR}(A_k) = \frac{n_{Rk} n_{Fk} n_{1k} n_{0k}}{T_k^2 (T_k - 1)}, \tag{2}$$

and

$$E(A_k) = \frac{n_{Rk} n_{1k}}{T_k} \tag{3}$$

where n_{Rk} and n_{Fk} represent the total number of reference and focal group members at ability level k, n_{1k} and n_{0k} represent the number of correct and incorrect responses at ability level k, and T_k equals the total number of examinees at ability level k. The MH chi-square is distributed approximately as a chi-square variate with one degree of freedom (Mantel & Haenszel, 1959).

Assessing DIF across multiple groups using the MH chi-square reduces to performing individual tests for each pair of groups to be compared, leading to the problem of an increased probability of committing a Type I error. Suppose that in the course of an analysis we wish to conduct several tests for DIF, each having an associated probability of a Type I error, α_i. Because the probability of committing a Type I error over repeated significance tests is larger than the probability on any one significance test, the probability of a Type I error over all tests exceeds the intended nominal alpha level. Using the terminology of Keppel (1991), the probability of committing a Type I error for a given comparison is referred to here as the error rate per comparison (α_i), and is distinguished from the probability of at least one Type I error across all tests of significance, referred to here as the familywise error rate (α_{FW}).

Two possible alternatives to the MH chi-square procedure to assess DIF across multiple groups are proposed here. The first solution is to adjust the per comparison alpha level (α_i) according to the Bonferroni inequality, which states that the

probability of making a Type I error anywhere in one of the *k* tests of significance is less than or equal to the sum of the Type I error rates of each test (Mendenhall, Scheaffer, & Wackerly, 1986). That is,

$$\alpha_{FW} \leq \sum_{i=1}^{j} \alpha_i \tag{4}$$

where, in the context of DIF detection across multiple focal groups, *i* refers to the test of DIF of the *i*th group of a total of *j* focal groups. If it is assumed that the probability of Type I error is equal for each of the *j* tests of significance, the familywise error rate expressed in Equation 4 can be approximated by

$$\alpha_{FW} \approx j\alpha_i \tag{5}$$

Using Equation 5, it is possible to determine the value of α_i required for each test to obtain a given value of α_{FW}. The adjusted alpha level for each test can be computed by

$$\alpha_i \approx \frac{\alpha_{FW}}{j} \tag{6}$$

where α_{FW} would be set equal to the intended nominal Type I error rate across all comparisons. Using the adjusted alpha level shown in Equation 6, the MH chi-square statistic can be performed for all *j* focal groups in relation to a single reference group, and the familywise error rate is guaranteed not to exceed the intended nominal Type I error rate per comparison when certain assumptions hold, such as equal reference and focal group ability distributions. To distinguish the MH procedure conducted with and without an adjusted value of α_i, the MH test performed with the Bonferroni-adjusted alpha level is denoted by BMH, whereas the MH test performed without the adjustment is denoted simply by MH.

Although the use of a Bonferroni-adjusted alpha level solves the problem of a spiraling Type I error rate, it still requires multiple tests of DIF. However, a natural extension of MH, the Generalized Mantel–Haenszel (GMH; Somes, 1986), can be used to test for DIF across all groups simultaneously. The GMH is considered to be a multivariate generalization of the MH chi-square statistic presented in Equation 1 (see Somes, 1986). Consider the data shown in Table 1 of correct and incorrect responses to a dichotomous item for *J* demographic groups.

The GMH test statistic is given by

$$\text{GMH}\chi^2 \approx \left[\mathbf{A}_k - E(\mathbf{A}_k)\right]' \left[\text{V}(A_k)\right]^{-1} \left[\mathbf{A}_k - E(\mathbf{A}_k)\right] \tag{7}$$

where, using the notation of Table 1

$$\mathbf{A}_k^{\downarrow} \therefore (n_{1Ak}, n_{1Bk}, ..., n_{1(J-1)k}) \tag{8}$$

$$E(\mathbf{A}_k^{\downarrow}) \therefore n_{1.k}\mathbf{n}_k^{\downarrow} / n_{..k} \tag{9}$$

$$\mathbf{n}_k^{\downarrow} \therefore (n_{.Ak}, n_{.Bk}, ..., n_{.(J-1)k}) \tag{10}$$

$$\mathrm{V}(\mathbf{A}_k) \therefore n_{1.k}n_{0.k} \frac{n_{..k}\mathrm{diag}(\mathbf{n}_k) - \mathbf{n}_k\mathbf{n}_k^{\downarrow}}{n_{..k}^2(n_{..k} - 1)} \tag{11}$$

where diag($\mathbf{n}_k$) is a (J–1)-by-(J–1) diagonal matrix with elements $\mathbf{n}_k$. Note that $\mathbf{A}_k$ and $E(\mathbf{A}_k)$ are vectors of length J–1, corresponding to any J–1 of the J demographic groups. The GMH statistic is distributed as a chi-square variable with J–1 degrees of freedom under the null hypothesis of no DIF. Note that the GMH procedure has already been described as a method of DIF detection by Zwick, Donoghue, and Grima (1993), whereby it was used to assess DIF between two demographic groups for polytomous items containing J possible nominal response categories. This previous application of the GMH procedure is analogous to that described here, with one exception: in the use made by Zwick et al. (1993), the group variable is dichotomous and the response variable is polytomous, whereas in the application presented here the group variable is polytomous and the response variable is dichotomous.

In this study I compare the performance of MH, BMH, and GMH under conditions in which several factors were varied, including the total number of focal groups, the number of focal groups experiencing DIF, the number of members in each group, the equality of the ability distributions of the reference and focal groups, and magnitude of matching criterion contamination. The power and Type I error rate of the three procedures are assessed using a simulation study.

TABLE 1
Data for the *k*th Level of the Ability Estimate for *J* Groups

	Group				
Item Score	*A*	*B*		*J*	*Total*
Correct (1)	n_{1Ak}	n_{1Bk}	—	n_{1Jk}	$n_{1.k}$
Incorrect (0)	n_{0Ak}	n_{0Bk}	—	n_{0Jk}	$n_{0.k}$
Total	$n_{.Ak}$	$n_{.Bk}$	—	$n_{.Jk}$	$n_{..k}$

METHOD

The simulations presented later were based on an artificial test of 60 dichotomous items. The parameters of the artificial items were determined according to a three-parameter logistic regression model (3PL). For each item, the difficulty parameter (*b*) was drawn from a normal distribution with a mean of zero and a standard deviation of one, and the item discrimination parameter (*a*) was sampled from a log-normal distribution where *a* is taken as the exponent of *z*, and *z* is a normal deviate with a mean of zero and a standard deviation of 0.1225. These parameter distributions are the same as those used in previous research (see Donoghue & Allen, 1993), and represent realistic distributions of item parameters. All items were assigned a *c*-parameter value of 0.2.

Generation of the simulated test data was conducted by drawing a standard normal variate (θ), computing the probability of success (*P*) on each item for each value of θ using the item's 3PL, drawing a uniform deviate (*U*) from the uniform distribution defined on (0, 1), and setting the item response equal to 0 if $U > P$ and 1 for $U \leq P$. DIF was introduced into one item on the artificial test by increasing the *b*-parameter by a constant 0.4 for the focal group only, making the item more difficult for the focal group relative to the reference group. The six factors examined in this study were (a) the number of groups being compared, (b) the number of focal groups experiencing DIF, (c) the number of members in each group, (d) equality of the means of the reference and focal group ability distributions, (e) the magnitude of DIF introduced into the studied item, and (f) the magnitude of matching criterion contamination. A discussion of these factors follows.

Factor 1

Four levels of the number of focal groups were considered: 1, 2, 3, and 4. Because there was always one common reference group to which each focal group was compared, the total number of groups was correspondingly 2, 3, 4, and 5. These four levels reflect a realistic range of the numbers of groups compared in practical DIF detection analyses. Note that in the case of only one focal group, MH and BMH are equivalent, and GMH is nearly identical to MH, the only difference being the correction for continuity of –½ expressed in the numerator of Equation 1.

Factor 2

Four levels of the number of focal groups experiencing DIF were considered. In the first level, only one of the focal groups contained DIF. For example, when there were a total of two focal groups, there was one focal group for which DIF was introduced into the scores, and one focal group for which DIF was not introduced into the scores. In the second, third, and fourth levels, there were two, three, and

four of the focal groups experiencing DIF. Note that the number of focal groups containing DIF was limited by the number of focal groups in the condition. That is, for the conditions having only two focal groups, the only possible levels of the number of focal groups containing DIF were 1 and 2. In contrast, for the conditions having four focal groups, their were four possible levels of the number of focal groups experiencing DIF.

Factor 3

Six levels of group size were considered. The first three levels assigned an equal number of members to each of the groups: 250, 500, and 1,000. The final three levels of group size were introduced to determine whether the detection of DIF in one focal group having a small number of members could be masked by the presence of other larger focal groups that do not display DIF. Thus, in the fourth level all focal groups experiencing DIF contained 250 members, whereas all other groups (the reference group and each focal group not experiencing DIF) contained 500 members. In the fifth level, all focal groups experiencing DIF contained 250 members, whereas all other groups contained 1,000. In the final level, all focal groups experiencing DIF contained 500 members, whereas all other groups contained 1,000 members.

Factor 4

Consideration was given to two levels of the mean of the focal and reference group ability distributions. The first level was a zero difference between the means of the reference and focal group ability distributions ($\mu_R = \mu_F = 0.0$). Because the reference group ability distribution was set to $N(0, 1)$, the focal group ability distribution was also $N(0, 1)$. The second level placed the mean of each focal group ability distribution one standard deviation below that of the reference group ($\mu_R = 0.0$, $\mu_F = -1.0$). Thus, in the second level, the focal group ability distribution was set to $N(-1, 1)$.

Factor 5

Two levels of the magnitude of DIF were considered: $b_R - b_F = 0.0$, and $b_R - b_F = 0.4$. The level $b_R - b_F = 0.0$ was used to assess the Type I error rate of the three procedures, and the level $b_R - b_F = 0.4$ was used to compare the power of the three procedures. The value of $b_R - b_F = 0.4$ was selected to represent a magnitude of DIF commonly found in applied testing situations, and is consistent with magnitudes used in previous DIF simulation research (Clauser, Mazor, & Hambleton, 1993; Donoghue & Allen, 1993; Donoghue, Holland, & Thayer, 1993). Although other

levels of introduced DIF could have been considered (e.g., b_R–b_F = 0.6), it is assumed that results from such conditions would offer little or no unique information concerning the relative performance of the three procedures. As a consequence, only the one level of b_R–b_F = 0.4 was used to compare the power of the three procedures.

Factor 6

Three levels of matching criterion contamination were considered in which 2, 5, and 10 items, other than the studied item, functioned differentially for certain focal groups. The groups experiencing contamination were either (a) all focal groups experiencing DIF on the studied item or (b) all focal groups not experiencing DIF on the studied item. DIF was introduced into all contaminating items by increasing the difficulty parameter for the focal groups experiencing contamination by 0.4 relative to the reference group (the same magnitude as the DIF introduced into the studied item). These magnitudes of contamination are consistent with those used in previous research investigating the effects of matching criterion contamination on MH (Clauser et al., 1993; Donoghue et al., 1993; Penfield, 2000).

For each condition, 1,000 trials were run. The performance of MH, BMH, and GMH were compared at each condition by recording the probability of detecting a statistically significant level of DIF for any one or more of the focal groups. For MH and GMH, the alpha level was set to 0.05 for all tests. The per comparison alpha level (α_i) used for BMH varied depending on the number of focal groups in the condition. Under the conditions of two, three, and four focal groups, the significance level for BMH was set to 0.05/2 = 0.025, 0.05/3 = 0.0167, and 0.05/4 = 0.0125, respectively.

RESULTS

The discussion of the results is organized into six sections investigating the following topics: (a) Type I error rate, (b) number of focal groups experiencing DIF, (c) overall group size, (d) unequal group sizes, (e) equality of reference and focal group ability distributions, and (f) matching criterion contamination. Unless otherwise stated, all results refer to the conditions in which there is no matching criterion contamination.

Type I Error Rate

Table 2 displays the Type I error rates for MH, BMH, and GMH across varying levels of group size, and total number of focal groups. Note that although GMH conducts a single test of significance for DIF among all groups, MH and BMH con-

TABLE 2
Type I Error Rate for MH, BMH, and GMH

	N(0, 1) Focal Group			*N(–1, 1) Focal Group*		
	MH	*BMH*	*GMH*	*MH*	*BMH*	*GMH*
One focal group						
N = 250	.03	—	—	.05	—	—
N = 500	.04	—	—	.04	—	—
N = 1,000	.05	—	—	.06	—	—
Two focal groups						
N = 250	.08	.04	.05	.07	.03	.05
N = 500	.06	.03	.05	.09	.05	.06
N = 1,000	.08	.05	.05	.10	.06	.06
Three focal groups						
N = 250	.10	.03	.05	.11	.04	.06
N = 500	.11	.03	.04	.12	.04	.06
N = 1,000	.12	.05	.06	.15	.06	.06
Four focal groups						
N = 250	.13	.03	.06	.13	.04	.06
N = 500	.11	.03	.04	.14	.04	.06
N = 1,000	.16	.04	.06	.18	.05	.06

Note. MH = Mantel–Haenszel procedure; BMH = MH procedure with a Bonferroni adjusted alpha level; GMH = generalized MH procedure.

duct an individual test of significance for each focal group. Thus, for MH and BMH, familywise Type I error rate is the probability of obtaining a significant result for any one of the focal groups involved in the analysis. In addition, note that for the case of one focal group, no results are reported for BMH and GMH. In this case, BMH is exactly equivalent to MH, and GMH is essentially equivalent to MH, with the exception of the –½ continuity correction in the numerator of MH (see Equation 1). Results have been reported for only those conditions in which all groups have the same number of members. Although other conditions were run in which the focal groups experiencing DIF had fewer members than the other groups, it was not expected that the Type I error rate would be affected by unequal focal groups sizes, and thus such results would offer no information concerning Type I error rate over that offered by the conditions in which all groups had equal sizes.

Consider first the Type I error rates when the ability distributions were equal for the reference and focal groups, displayed in the left-hand side of Table 2. The familywise Type I error rate for BMH was consistently at or below 0.05 across all conditions. Similarly, the Type I error rate of GMH was consistently at the nominal level of 0.05. In contrast to BMH and GMH, the familywise Type I error rate of MH consistently exceeded the nominal per-comparison level of 0.05 when there

were two or more focal groups. The extent to which MH displayed an inflated Type I error rate increased with the total number of focal groups tested and the size of the groups, reaching 0.16 in the condition of four focal groups, each having 1,000 members.

The right-hand side of Table 2 displays the Type I error rates under the conditions in which the mean of each focal group ability distribution was set to one standard deviation below that of the reference group. In this case, both BMH and GMH displayed only a slight inflation in Type I error rate to 0.06, primarily when group sizes were large. Paralleling the results obtained when the ability distributions were equal for reference and focal groups, the Type I error rate of MH far exceeded the nominal per-comparison level, reaching 0.18 under the condition of four focal groups of 1000 members each. Note that the slight increase in Type I error rate observed across all three procedures when the mean of the reference and focal group ability distributions differ is consistent with the results of previous theoretical (Holland & Thayer, 1988; Zwick, 1990) and empirical (Clauser et al., 1993; Penfield, 1999, 2000) research of the MH procedure in the two-group case.

The grossly inflated Type I error rates of MH are not surprising, given the number of comparisons made and the differences in the reference and focal group ability distributions. Similarly, the subnominal Type I error rates of BMH were expected given that the nominal per-comparison alpha level is the upper bound to the Bonferroni adjusted alpha level when reference and focal group ability distributions are equal. What is of interest here is that the Type I error rate of GMH remained close to the nominal level of 0.05, even when there were differences in the group ability distributions, suggesting that the use of GMH adequately controlled the inflated Type I error rates experienced by two-group DIF detection methods such as MH.

Number of Groups Experiencing DIF

The results of the power of MH, BMH, and GMH are displayed in Tables 3 and 4 for the conditions in which the focal group ability distribution was set to $N(0, 1)$ and $N(-1, 1)$, respectively. Each row in Tables 3 and 4 corresponds to a particular combination of the total number of focal groups (from one to four) and the number of focal groups experiencing DIF (from one to all). Each column represents a different level of group size; the first three columns represent conditions in which all groups had the same size (250, 500, and 1,000), and the last three columns represent conditions in which the focal groups experiencing DIF had fewer members than the other groups (250 vs. 500, 250 vs. 1,000, and 500 vs. 1,000).

Tables 3 and 4 present the results of all conditions, and are included here to permit the reader to explore the entire set of results. However, the large amount of information presented in Tables 3 and 4 precludes an efficient comparison of the performance of the three methods. To facilitate an assessment of the most intriguing

TABLE 3
Power When Focal Group Ability Distributions Are *N*(0, 1)

		Group Sizes					
		250	*500*	*1,000*	*500(250)*	*1,000(250)*	*1,000(500)*
One focal group							
No. DIF = 1	MH	.44	.70	.90	.53	.65	.80
Two focal groups							
No. DIF = 1	MH	.44	.73	.90	.57	.65	.83
	BMH	.33	.64	.85	.46	.56	.77
	GMH	.49	.76	.90	.58	.63	.82
No. DIF = 2	MH	.60	.82	.95	.72	.80	.89
	BMH	.49	.76	.92	.63	.74	.86
	GMH	.50	.76	.92	.67	.77	.87
Three focal groups							
No. DIF = 1	MH	.48	.74	.92	.61	.65	.83
	BMH	.33	.61	.86	.43	.50	.74
	GMH	.55	.77	.93	.58	.59	.81
No. DIF = 2	MH	.60	.84	.94	.75	.87	.91
	BMH	.42	.72	.90	.60	.72	.83
	GMH	.61	.85	.94	.73	.82	.89
No. DIF = 3	MH	.66	.88	.95	.81	.87	.91
	BMH	.50	.79	.92	.66	.77	.88
	GMH	.48	.77	.91	.69	.81	.88
Four focal groups							
No. DIF = 1	MH	.48	.74	.91	.60	.69	.82
	BMH	.25	.57	.83	.38	.50	.69
	GMH	.48	.76	.91	.51	.56	.79
No. DIF = 2	MH	.59	.84	.95	.77	.80	.93
	BMH	.40	.72	.91	.57	.66	.82
	GMH	.64	.86	.95	.75	.79	.91
No. DIF = 3	MH	.66	.88	.95	.82	.89	.93
	BMH	.47	.78	.90	.64	.74	.85
	GMH	.64	.88	.94	.79	.85	.91
No. DIF = 4	MH	.72	.91	.96	.85	.90	.96
	BMH	.52	.80	.93	.68	.79	.91
	GMH	.47	.76	.91	.68	.82	.90

Note. Group sizes shown in parentheses indicate the size of the focal groups experiencing DIF. DIF = differential item functioning; MH = Mantel–Haenszel procedure; BMH = MH procedure with a Bonferroni adjusted alpha level; GMH = generalized MH procedure.

comparisons—GMH versus BMH, and GMH versus MH—two ratios were computed for each condition: (a) the power of GMH divided by the power of BMH, and (b) the power of GMH divided by the power of MH. Any value of these ratios exceeding unity indicates a relatively higher power of GMH, and any value of these ratios smaller than unity indicates a relatively lower power of GMH. These ratios

TABLE 4
Power When Focal Group Ability Distributions Are *N*(–1, 1)

		Group Sizes					
		250	*500*	*1,000*	*500(250)*	*1,000(250)*	*1,000(500)*
One focal group							
No. DIF = 1	MH	.31	.53	.71	.40	.45	.61
Two focal groups							
No. DIF = 1	MH	.32	.56	.76	.43	.49	.69
	BMH	.23	.48	.71	.34	.39	.62
	GMH	.32	.57	.73	.41	.45	.66
No. DIF = 2	MH	.43	.65	.82	.53	.61	.71
	BMH	.35	.59	.77	.44	.52	.65
	GMH	.38	.62	.81	.50	.58	.69
Three focal groups							
No. DIF = 1	MH	.38	.59	.76	.44	.48	.65
	BMH	.22	.44	.66	.30	.34	.52
	GMH	.33	.56	.71	.38	.39	.59
No. DIF = 2	MH	.48	.70	.82	.57	.67	.73
	BMH	.32	.58	.75	.43	.53	.63
	GMH	.45	.68	.81	.54	.61	.70
No. DIF = 3	MH	.51	.72	.82	.60	.67	.78
	BMH	.36	.60	.75	.46	.54	.69
	GMH	.38	.63	.79	.50	.62	.74
Four focal groups							
No. DIF = 1	MH	.40	.61	.79	.45	.53	.68
	BMH	.21	.44	.67	.27	.35	.53
	GMH	.34	.56	.72	.37	.40	.58
No. DIF = 2	MH	.48	.69	.84	.60	.66	.74
	BMH	.28	.53	.75	.38	.48	.60
	GMH	.45	.66	.84	.52	.59	.70
No. DIF = 3	MH	.51	.73	.84	.62	.68	.80
	BMH	.31	.59	.75	.44	.49	.67
	GMH	.45	.70	.82	.56	.62	.76
No. DIF = 4	MH	.57	.75	.85	.67	.71	.80
	BMH	.37	.61	.76	.49	.58	.68
	GMH	.38	.64	.81	.57	.64	.74

Note. Group sizes shown in parentheses indicate the size of the focal groups experiencing DIF. DIF = differential item functioning; MH = Mantel–Haenszel procedure; BMH = MH procedure with a Bonferroni adjusted alpha level; GMH = generalized MH procedure.

are displayed in Tables 5 and 6, and will be the basis of the discussion of the power of the three procedures presented here.

Consider first the comparison of power between GMH and BMH shown in Table 5. Examination of the power ratios indicates that the power of GMH relative to that of BMH was dependent on the number of focal groups experiencing DIF.

TABLE 5
Ratio of Power of GMH Over Power of BMH

	Group Sizes					
	250	*500*	*1,000*	*500(250)*	*1,000(250)*	*1,000(500)*
N(0, 1) focal group						
Two focal groups						
No. DIF = 1	1.48	1.19	1.06	1.26	1.13	1.06
No. DIF = 2	1.02	1.00	1.00	1.06	1.04	1.01
Three focal groups						
No. DIF = 1	1.67	1.26	1.08	1.35	1.18	1.09
No. DIF = 2	1.45	1.18	1.04	1.22	1.14	1.07
No. DIF = 3	0.96	0.97	0.99	1.05	1.05	1.00
Four focal groups						
No. DIF = 1	1.92	1.33	1.10	1.34	1.12	1.14
No. DIF = 2	1.60	1.19	1.04	1.32	1.20	1.11
No. DIF = 3	1.36	1.12	1.04	1.23	1.15	1.07
No. DIF = 4	0.90	0.95	0.98	1.00	1.04	0.99
N(–1, 1) focal group						
Two focal groups						
No. DIF = 1	1.39	1.19	1.03	1.21	1.15	1.13
No. DIF = 2	1.08	1.05	1.05	1.14	1.12	1.06
Three focal groups						
No. DIF = 1	1.50	1.27	1.08	1.27	1.15	1.13
No. DIF = 2	1.41	1.17	1.08	1.26	1.15	1.11
No. DIF = 3	1.06	1.05	1.05	1.09	1.15	1.07
Four focal groups						
No. DIF = 1	1.62	1.27	1.07	1.37	1.14	1.09
No. DIF = 2	1.61	1.25	1.12	1.37	1.23	1.17
No. DIF = 3	1.45	1.19	1.09	1.27	1.27	1.13
No. DIF = 4	1.03	1.05	1.07	1.16	1.10	1.09

Note. Group sizes shown in parentheses indicate the size of the focal groups experiencing DIF. DIF = differential item functioning; MH = Mantel–Haenszel procedure; BMH = MH procedure with a Bonferroni adjusted alpha level; GMH = generalized MH procedure.

When only one focal group experienced DIF (e.g., one of two focal groups, or one of three focal groups) the power of GMH was consistently higher than that of BMH, a result that was consistent for both levels of focal group ability distribution. The extent to which the power of GMH exceeded that of BMH was dependent on both the number of members in each group, and the total number of focal groups involved. When group sizes were small, the power of GMH ranged between 1.39 and 1.92 times that of BMH. As groups sizes increased to 1,000, the power ratio decreased to values just slightly greater than unity. A similar result was obtained when a portion, but not all, of the focal groups experienced DIF (e.g., two of three focal groups). However, this same trend was not observed when all focal groups

TABLE 6
Ratio of Power of GMH Over Power of MH

	Group Sizes					
	250	*500*	*1,000*	*500(250)*	*1,000(250)*	*1,000(500)*
N(0, 1) focal group						
Two focal groups						
No. DIF = 1	1.19	1.04	1.00	1.02	0.97	0.99
No. DIF = 2	0.83	0.93	0.97	0.93	0.96	0.98
Three focal groups						
No. DIF = 1	1.15	1.04	1.01	0.95	0.91	0.98
No. DIF = 2	1.02	1.01	1.00	0.97	0.94	0.98
No. DIF = 3	0.73	0.88	0.96	0.85	0.93	0.97
Four focal groups						
No. DIF = 1	1.00	1.03	1.00	0.85	0.81	0.96
No. DIF = 2	1.08	1.02	1.00	0.97	0.99	0.98
No. DIF = 3	0.97	1.00	0.99	0.96	0.96	0.98
No. DIF = 4	0.65	0.84	0.95	0.80	0.91	0.94
N(–1, 1) focal group						
Two focal groups						
No. DIF = 1	1.00	1.02	0.96	0.95	0.92	0.96
No. DIF = 2	0.88	0.95	0.99	0.94	0.95	0.97
Three focal groups						
No. DIF = 1	0.87	0.95	0.93	0.86	0.81	0.91
No. DIF = 2	0.94	0.97	0.99	0.95	0.91	0.96
No. DIF = 3	0.75	0.88	0.96	0.83	0.93	0.95
Four focal groups						
No. DIF = 1	0.85	0.92	0.91	0.82	0.75	0.85
No. DIF = 2	0.94	0.96	1.00	0.87	0.89	0.95
No. DIF = 3	0.88	0.96	0.98	0.90	0.91	0.95
No. DIF = 4	0.67	0.85	0.96	0.85	0.90	0.93

Note. Group sizes shown in parentheses indicate the size of the focal groups experiencing DIF. DIF = differential item functioning; MH = Mantel–Haenszel procedure; GMH = generalized MH procedure.

experienced DIF. In this case, the power of GMH was less than that of BMH when the focal group ability distribution was set to *N*(0, 1), and only slightly greater than that of BMH when the focal group ability distribution was set to *N*(–1, 1).

Next, consider the comparison of the power of GMH to MH shown in Table 6. When only one of the focal groups experienced DIF, the power of GMH slightly exceeded that of MH when the focal group ability distribution was set to *N*(0, 1). Note that the superior power of GMH tended to disappear when the focal group experiencing DIF had fewer members than the other groups. When only one focal group experienced DIF, and the focal group ability distribution was set to *N*(–1, 1), the power ratio of GMH over MH was consistently less than unity, indicating

higher powers for MH. This result was expected given the increased Type I error rate of MH under this condition (see Table 2). The higher relative power of MH was accentuated when the focal group experiencing DIF had fewer members than the other groups. Similar results were obtained when more than one, but fewer than all, focal groups experienced DIF. When all focal groups experienced DIF, the power ratios indicated a very poor power of GMH relative to MH, reaching as low as 0.65. Interestingly, this poor relative performance of GMH when all groups had equal sizes was moderated when the focal groups experiencing DIF (all focal groups in this case) had fewer members than the reference group.

Overall Group Size

It was of interest to examine how the power of MH, BMH, and GMH compared as groups sizes ranged from small ($N = 250$) to large ($N = 1{,}000$). The first three columns of Tables 3 and 4 display the power of the statistics for the conditions in which all groups had the same number of members. Three results are worth mentioning. First, as expected, the power increased for all three procedures as the group size increased. Under the small group sizes, power typically ranged between 0.30 to 0.40 when only one focal group experienced DIF and between 0.40 and 0.70 when all focal groups experienced DIF. These powers increased dramatically to over 0.90 as group sizes increased to $N = 1{,}000$.

The second result of interest is that the differences in the power of the three procedures decreased as group size increased. Under the small size, differences in the power among the three procedures was often over 0.20. However, when group sizes were large, differences were generally no greater than 0.05. This is not surprising, because at the larger group sizes the power of the three procedures was approaching the theoretical limit of 1.0, causing a ceiling effect on the observed powers.

The third result of interest is that, in general, the relative performance of the three procedures was not substantially affected by group size. Although the differences in the power between the three procedures decreased as group size increased, the same general ordering of power remained the same.

Unequal Group Sizes

The extent to which unequal group sizes affected the power of the three procedures has already been touched on in previous sections. This section serves to elaborate on specific trends observed in the results. It was decided to examine the effects of unequal focal group sizes by comparing the power observed when all groups had 1,000 members to the power observed when groups experiencing DIF had 250 members and the groups not experiencing DIF had 1,000 members. This comparison was made by taking the ratio of the power obtained in the unequal group size

case (250 vs. 1,000 members) over that obtained in the equal group size case (all groups had 1,000 members). The extent to which these power ratios are less than unity serves as a measure of how affected the procedure was by decreasing the size of the focal groups experiencing DIF. Only the conditions in which the focal groups experiencing DIF had 250 members were used here for comparison because this was the most extreme case, and thus offered the greatest potential for uncovering trends in the data. Table 7 displays these power ratios over varying levels of the total number of focal groups, the number of focal groups experiencing DIF, and focal group ability distributions.

Three patterns emerge from the power ratios displayed in Table 7. First, as the number of focal groups experiencing DIF increased, the power ratios tended to increase, indicating less of an effect of differential focal group sizes. In general, the power ratios for the conditions in which only one focal group experienced DIF were substantially lower than the power ratios observed in other conditions. This finding was consistent for all three procedures, and across both levels of the focal group ability distribution. Second, MH tended to have the highest power ratios, followed by GMH, and then BMH. This suggests that BMH was the most severely affected procedure by differential focal group sizes. Third, the power ratios were larger when the focal group ability distributions were equal to that of the reference

TABLE 7
Effects of Differential Focal Group Sizes on the Power of MH, BMH, and GMH

	N(0, 1) Focal Group			*N(–1, 1) Focal Group*		
	MH	*BMH*	*GMH*	*MH*	*BMH*	*GMH*
One focal group						
No. DIF = 1	.72	—	—	.63	—	—
Two focal groups						
No. DIF = 1	.72	.66	.70	.64	.55	.62
No. DIF = 2	.84	.80	.84	.74	.68	.72
Three focal groups						
No. DIF = 1	.71	.58	.63	.63	.52	.55
No. DIF = 2	.93	.80	.87	.82	.71	.75
No. DIF = 3	.92	.84	.89	.82	.72	.78
Four focal groups						
No. DIF = 1	.76	.76	.62	.67	.52	.56
No. DIF = 2	.84	.73	.83	.79	.64	.70
No. DIF = 3	.94	.82	.90	.81	.65	.76
No. DIF = 4	.94	.85	.90	.84	.76	.79

Note. The values shown represent the ratio of the power obtained when the focal groups experiencing DIF had 250 members and all other groups had 1,000 members over the power obtained when all groups had 1,000 members. DIF = differential item functioning; MH = Mantel–Haenszel procedure; BMH = MH procedure with a Bonferroni adjusted alpha level; GMH = generalized MH procedure.

group, indicating that the effects of differential focal group sizes were most severe when the focal groups had ability distributions with means below that of the reference group.

Equality of Ability Distributions

The effect of the equality of focal and reference group ability distributions has already been investigated in relation to several of the factors considered in the simulation. What remains to be investigated is how the overall power of the three procedures is affected by the equality of the reference and focal group ability distributions. Tables 3 and 4 indicate that the power of all three procedures were consistently lower in the conditions for which the focal group ability distributions were set to $N(-1, 1)$. This result was expected because when the focal group ability distribution is set to $N(-1, 1)$, the majority of the focal group members lie in a region of the ability continuum (the lower portion of the continuum) for which there is little difference in the expected performance of the reference and focal groups. This effect is consistent with that observed by Penfield (1999) studying the effect of overall sample mean ability on the power of MH in the two-group case.

How does a difference in the mean of the reference and focal group ability distributions affect the relative performance of the three procedures? To address this question, for each condition a ratio was computed by dividing the power obtained in the $N(-1, 1)$ condition by that obtained in the $N(0, 1)$ condition. These results (not presented here) suggest that all three methods were approximately equally affected by unequal group ability distributions. Furthermore, there was no consistent relation of the effect of equality of group ability distributions and the number of focal groups, the number of focal groups experiencing DIF, or equality of focal group sizes. The only consistent result observed was that the effect of unequal group ability distributions decreased as group sizes increased.

Matching Criterion Contamination

The results concerning the effect of matching criterion contamination on the Type I error rate and power of MH, BMH, and GMH for the conditions in which there were a total of three focal groups are displayed in Tables 8 and 9, respectively. The results reported here are restricted to conditions in which all groups had 500 members. Although only the conditions in which there were three focal groups having 500 members each are shown here, these results are representative of those observed across other levels of number of focal groups and group size. Note that the top half of the results displayed in Tables 8 and 9 represents the conditions in which the focal groups experiencing contamination were the same as those experiencing DIF for the studied item, whereas the bottom half of the table represents the

TABLE 8
Effects of Contamination on the Type I Error Rate of MH, BMH, and GMH

		Number of Contaminating Items					
		N(0, 1) Focal Group			*N(–1, 1) Focal Group*		
		2	*5*	*10*	*2*	*5*	*10*
Same Groups							
No. DIF = 1	MH	.09	.10	.15	.10	.11	.16
	BMH	.03	.03	.06	.04	.05	.05
	GMH	.05	.06	.11	.05	.07	.08
No. DIF = 2	MH	.12	.12	.17	.11	.12	.15
	BMH	.05	.05	.06	.04	.04	.05
	GMH	.06	.07	.10	.04	.06	.09
No. DIF = 3	MH	.11	.10	.18	.12	.11	.17
	BMH	.04	.05	.08	.05	.05	.06
	GMH	.05	.05	.09	.05	.06	.07
Other Groups							
No. DIF = 1	MH	.10	.13	.16	.13	.12	.17
	BMH	.02	.05	.06	.05	.05	.06
	GMH	.04	.08	.09	.06	.06	.08
No. DIF = 2	MH	.11	.11	.13	.12	.12	.13
	BMH	.04	.04	.04	.04	.04	.04
	GMH	.06	.06	.09	.06	.06	.08

Note. Same Groups indicates that the focal groups experiencing contamination were the same as those experiencing DIF in the studied item, whereas Other Groups indicates that the focal groups experiencing contamination were those focal groups not experiencing DIF for the studied item. Note that these results correspond to the conditions in which there were three focal groups, and all groups had 500 members. DIF = differential item functioning; MH = Mantel–Haenszel procedure; BMH = MH procedure with a Bonferroni adjusted alpha level; GMH = generalized MH procedure.

results obtained when the focal groups experiencing contamination were those groups not experiencing DIF in the studied item.

Consider first the effects of contamination on Type I error rate, shown in Table 8. The results indicate that as contamination increased, the Type I error rates increased. This result is consistent with that found in previous research (Penfield, 2000). Although this result was observed for all three procedures, it was strongest for MH, and weakest for BMH. This result was observed across both levels of focal group ability distribution and for all levels of the number of groups experiencing contamination.

Next, consider the effects of contamination on power, displayed in Table 9. In general, for all three methods, power decreased as contamination increased, a result that is consistent with the findings of previous research (Penfield, 2000). Examining the relative effects of contamination on the three procedures, we see that there was little differential effect of contamination on the power of MH, BMH, and

TABLE 9
Effects of Contamination on the Power of MH, BMH, and GMH

		Number of Contaminating Items					
		N(0, 1) Focal Group			*N(–1, 1) Focal Group*		
		2	*5*	*10*	*2*	*5*	*10*
Same Groups							
No. DIF = 1	MH	.72	.67	.56	.57	.52	.44
	BMH	.56	.50	.39	.42	.36	.27
	GMH	.73	.70	.57	.52	.49	.40
No. DIF = 2	MH	.83	.79	.70	.65	.59	.56
	BMH	.73	.66	.51	.53	.44	.38
	GMH	.83	.81	.70	.64	.56	.53
No. DIF = 3	MH	.87	.81	.73	.72	.66	.59
	BMH	.75	.70	.60	.61	.52	.42
	GMH	.74	.68	.56	.64	.56	.45
Other Groups							
No. DIF = 1	MH	.79	.77	.76	.59	.59	.64
	BMH	.65	.62	.62	.44	.44	.49
	GMH	.85	.83	.85	.55	.58	.63
No. DIF = 2	MH	.85	.84	.85	.68	.68	.69
	BMH	.76	.74	.72	.54	.55	.54
	GMH	.88	.86	.87	.66	.66	.66

Note. Same Groups indicates that the focal groups experiencing contamination were the same as those experiencing DIF in the studied item, whereas Other Groups indicates that the focal groups experiencing contamination were those focal groups not experiencing DIF for the studied item. Note that these results correspond to the conditions in which there were three focal groups, and all groups had 500 members. DIF = differential item functioning; MH = Mantel–Haenszel procedure; BMH = MH procedure with a Bonferroni adjusted alpha level; GMH = generalized MH procedure.

GMH when the groups experiencing contamination were the same as those experiencing DIF in the studied item. However, when the focal groups experiencing contamination were those groups not experiencing DIF for the studied item, the performance of GMH increased substantially relative to the other procedures. This result was consistent across both levels of focal group ability distribution.

DISCUSSION

In this study I compared the performance of three DIF detection procedures in assessing DIF among multiple groups: (a) the MH, (b) the BMH, and (c) the GMH. Several general results were observed. First, the Type I error rate of MH reached unacceptably high levels as the number of groups increased, whereas that of BMH and GMH remained at the nominal level. Second, the power of

BMH was consistently substantially lower than that of both MH and GMH. The power of GMH and MH were similar over most conditions, with the exception of conditions in which all focal groups experienced DIF, in which case the power of GMH fell dramatically lower than that of MH. Third, the magnitude of the difference in power of the three procedures was dependent on the size of the sample, being largest when groups had small sizes (N = 250), and diminishing substantially as groups reached their largest size (N = 1,000). This finding suggests that the differential performance of the three procedures is most apparent when sample sizes are small. Fourth, although all three procedures exhibited a decrease in power when the focal groups experiencing DIF had fewer members than the other groups, this decrease in power was most apparent for BMH, followed by GMH, and then MH. Fifth, under certain conditions of matching criterion contamination, GMH displayed power rates substantially higher than those of MH.

A primary aim of this study was to conduct a comprehensive comparison of the performance of MH, BMH, and GMH to shed light on which procedure should be used, and under which conditions the conclusions hold. The results suggest that GMH is the most appropriate statistical procedure to be used when DIF is being assessed across multiple focal groups. This conclusion is based on two results. First, although MH displayed Type I error rates that spiraled to unacceptably high levels as the number of groups increased, GMH maintained a Type I error rate very close to the nominal level across all conditions. Second, the power of GMH was generally equal to or greater than that of BMH, and was similar to that of MH, so long as not all focal groups experienced DIF. When the focal and reference groups had equal ability distributions, it was frequently the case that GMH displayed a power exceeding that of MH, even though MH displayed higher Type I error rates for the very same conditions. This was particularly the case when only one of the focal groups experienced DIF. These results indicate the clear superior performance of GMH over both MH and BMH when multiple focal groups are being assessed for DIF.

Aside from displaying an overall superior performance in terms of power and Type I error rate, there are several other important advantages to using GMH. First, GMH offers a simpler method for assessing DIF among multiple groups. Test developers concerned with the detection of bias are often faced with the need to conduct DIF analyses at all stages of test development, including pilot, field, and operational test administrations. The results presented here indicate that GMH is a simple and efficient screening mechanism for DIF because all groups can be tested simultaneously. Should a nonsignificant result be obtained, then it has been shown with a certain level of confidence that DIF does not exist among the groups considered. In this case, using GMH in place of individual tests for each group has potentially saved time and resources. If a significant value of GMH is obtained, then the item has been shown to function differently for two or more groups. To determine

which groups are experiencing DIF, post hoc paired comparisons can be performed between each focal group and the reference group using the BMH procedure, thus ensuring that the Type I error rate across all comparisons does not exceed the intended nominal familywise error level. This recommendation is consistent with that made by Kim et al. (1995), whereby on observing a significant value of the Q_j statistic, the authors recommended using Lord's chi-square procedure with an adjusted alpha level to assess the magnitude of DIF between any two of the groups.

An additional advantage of GMH is that it can be used to test for DIF between any pair of J definable groups. In this situation, the number of pairwise comparisons increases dramatically as the number of groups increases. Under such circumstances, the MH Type I error rate can be expected to increase dramatically as the number of groups increases, and the power of MHB can be expected to decrease substantially as the number of groups increases. Because GMH consists of a single test of significance regardless of the number of groups compared, it holds promise to perform the best under these circumstances. In addition, as previously discussed, GMH would permit an assessment of DIF among all J groups with a single test, thus offering an efficient method relative to the standard two-group DIF detection methods.

Although the results of this study indicate that GMH is the most effective method for assessing DIF among multiple groups, there are several limitations of these results that deserve recognition. First, in this study I considered only a consistent magnitude of DIF, set by increasing the difficulty parameter of the studied item by 0.4 for all focal groups experiencing DIF. The extent to which the comparative performance of the three methods would change if different levels of DIF were introduced into different focal groups is not addressed in this study.

A second limitation of this study is that it does not offer information on how GMH would be effectively employed in cases where sample sizes become very large. A problem encountered using a chi-square test of DIF is that when sample sizes become large (e.g., $N > 1{,}000$), statistical significance is often observed even though no substantially meaningful level of DIF is found in the data. To address this concern in the two-group case, the magnitude of DIF in an item is typically assessed using a combination of statistical significance and effect size (Zieky, 1993). In the case of MH, the Mantel–Haenszel common odds ratio is transformed to the symmetrical *MHD–DIF* index, which has a mean of zero and a known standard error (Philips & Holland, 1987; Robins, Breslow, & Greenland, 1986). Although the exact sampling distribution of the *MHD–DIF* index is unknown, its standard error can be used to obtain an approximate confidence interval. Using the absolute value of the *MHD–DIF* index as a measure of DIF effect size, along with the results of statistical significance from zero and unity, it is possible to classify the amount of DIF in the item as negligible, slight to moderate, and moderate to high (Zieky, 1993). This study offers no information on how such a DIF assessment strategy

could be applied to GMH. At present, no known measure of effect size associated with GMH exists, and thus no method for assessing multiple group DIF that incorporates both statistical significance and effect size is available. It should be noted that measures of effect size for chi-square statistics of single contingency table data do exist (Cohen, 1977, pp. 216–227). The development of a suitable measure of effect size of DIF across all groups simultaneously could be accomplished by obtaining a composite of effect sizes across all *m* response-by-group contingency tables (i.e., Table 1), where *m* equals the number of matching categories used to group reference and focal group members of equal estimated ability. Further research is required to develop such a measure of effect size for multiple group DIF and determine its performance in quantifying the magnitude of DIF across multiple groups.

The previous discussion of the limitations of this study highlighted several questions which are left for future research. In addition to these questions, several other lines of inquiry should be considered for future investigation. First, how does the performance of the Q_j statistic proposed by Kim et al. (1995) compare to that of GMH? This question could be answered using a simulation study similar to the one employed here to permit a comparison of the two procedures across a variety of conditions. Second, how does GMH perform when DIF is nonuniform? It has been shown that when DIF is nonuniform, the power of the MH chi-square statistic is substantially reduced, causing other procedures such as logistic regression to be preferred (Swaminathan & Rogers, 1990). Because logistic regression models that include additional independent variables coding for multiple groups are a natural extension of the logistic regression model used for the two-group case, such procedures may be an effective alternative for assessing DIF among multiple groups when the DIF is nonuniform. Logistic regression procedures have also been developed for assessing bias in polytomously scored items (French & Miller, 1996), and thus offer a method of assessing DIF in performance assessments across multiple groups.

REFERENCES

Angoff, W. H. (1972, September). *A technique for the investigation of cultural differences.* Paper presented at the annual meeting of the American Psychological Association, Honolulu, HI. (ERIC Document Reproduction Service No. ED 069 686)

Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement, 34,* 807–816.

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

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–44.

Clauser, B. E., Mazor, K., & 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.

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28*, 61–75.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 25–29). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Donoghue, J. R., & Allen, N. L. (1993). Thin vs. 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 factors that affect the Mantel–Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137–166). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77*, 177–184.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differentially item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315–332.

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

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249–260.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Kim, S. -H, Cohen, A. H., & Park, T. -H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*, 261–276.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349–364). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Amsterdam, The Netherlands: Swets and Zeitlinger B. V.

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

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.

Mendenhall, W., Scheaffer, R. L., & Wackerly, D. D. (1986). *Mathematical statistics with applications* (3rd ed.). Boston: Duxbury.

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

Penfield, R. D. (1999, October). *The effects of sample ability on DIF detection.* Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, NY.

Penfield, R. D. (2000). *The effects of matching criterion contamination on the Mantel–Haenszel procedure.* Unpublished doctoral dissertation, Department of Curriculum, Teaching and Learning, University of Toronto, Canada.

Penfield, R. D., & Lam, T. C. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5–15.

Philips, A., & Holland, P. W. (1987). Estimators of the variance of the Mantel–Haenszel log-odds-ratio estimate. *Biometrics, 43,* 425–431.

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 and large-strata limiting models. *Biometrics, 42,* 311–323.

Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement, 25,* 1–13.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27,* 67–81.

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

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

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

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.