The Alberta Journal of Educational Research *Vol. XLIX, No. 3, Fall 2003, 231-243*

Randall D. Penfield
University of Florida

Applying the Breslow-Day Test of Trend in Odds Ratio Heterogeneity to the Analysis of Nonuniform DIF

This article applies the Breslow-Day test of trend in odds ratio heterogeneity (BD) to the detection of nonuniform DIF. A simulation study was conducted to assess the power and Type I error rate of BD, as well as a combined decision rule (CDR) whereby a decision of the existence of DIF was based on a combination of the decisions made using BD and the Mantel-Haenszel chi-square. The results indicated that CDR displayed good Type I error rate and power across a variety of conditions. Comparing these results with those of earlier research indicates that CDR may yield more accurate decisions about DIF than other commonly used DIF detection procedures.

Item bias is an important consideration when assessing the validity of achievement tests. Typically, the presence of item bias is assessed using the framework of differential item functioning (DIF), defined as a difference in measurement properties of an item for two groups (Camilli & Shepard, 1994; Dorans & Holland, 1993; Hanson, 1998). By convention, the item under investigation is referred to as the studied item, and the groups being compared are referred to as the reference and focal groups. The existence of DIF can be assessed using a variety of statistical procedures, including item response theory (Lord, 1980), logistic regression (Swaminathan & Rogers, 1990), and contingency-table methods (Camilli & Shepard, 1994; Dorans & Holland, 1993; Holland & Thayer, 1988). Descriptions of DIF detection procedures for dichotomous items are provided by Camilli and Shepard (1994), Clauser and Mazor (1998), Hills (1989), and Millsap and Everson (1993).

An important distinction to be made when conceptualizing DIF is that between uniform and nonuniform DIF. Uniform DIF exists when the level of DIF is independent of ability level. If the level of DIF is quantified using the odds ratio (the ratio of odds of correct response for the reference group over that of the focal group), then uniform DIF exists when the odds ratio is not equal to unity, but remains constant across the ability continuum. In contrast, nonuniform DIF exists when the odds ratio varies systematically across the ability continuum. Under certain conditions, nonuniform DIF may lead to the situation whereby one group displays a strong relative advantage at one end of the ability continuum, and the second group displays a strong relative advantage at the opposite end of the ability continuum. This form of nonuniform DIF is typically referred to as crossing-nonuniform DIF. In the context of item response theory (IRT), uniform DIF exists when there is a between-group

---

Randall Penfield is an assistant professor in the Department of Educational Psychology. He specializes in educational measurement and psychometrics.

difference in the b parameters only, nonuniform DIF exists when there is a between-group difference in the a parameters (regardless of any between-group difference in the b parameters), and crossing-nonuniform DIF exists when there is a between-group difference in the a parameters but no substantial difference in the b parameters.

Determining whether DIF is nonuniform, and in particular crossing-nonuniform, is an important step in DIF analyses. This importance stems from the fact that many of the commonly employed DIF detection procedures work under the assumption that the DIF is uniform; for example, the Mantel-Haenszel common odds ratio (Dorans & Holland, 1993; Holland & Thayer, 1993; Mantel & Haenszel, 1959), the Mantel-Haenszel chi-square (Camilli & Shepard, 1994; Mantel & Haenszel, 1959), SIBTEST (Shealy & Stout, 1993), the standardized p-difference (Dorans & Kulick, 1986), and ANOVA-based methods (Whitmore & Schumacker, 1999). As a result, many of the most popular methods for detecting DIF have unacceptably low power for detecting nonuniform DIF. For example, Swaminathan and Rogers (1990) showed the Mantel-Haenszel procedure to be completely ineffective at detecting crossing-nonuniform DIF, having a power equal to the nominal Type I error rate. Although uniform DIF is observed more frequently than nonuniform DIF, nonuniform DIF has been identified in applied DIF analyses (Hambleton & Rogers, 1989), and thus is a possible threat to item validity.

Several procedures have been developed for detecting nonuniform DIF, including logistic regression (Narayanan & Swaminathan, 1996; Swaminathan & Rogers, 1990), crossing SIBTEST (Li & Stout, 1993), and a split Mantel-Haenszel procedure (Mazor, Clauser, & Hambleton, 1994). Although these procedures display good power for detecting nonuniform DIF, they have several disadvantages that hamper their practical utility. The most severe disadvantage is that of an inflated Type I error rate; that is, the logistic regression procedure has displayed a Type I error rate that was on the order of two (Narayanan & Swaminathan, 1996) and seven (Whitmore & Schumacker, 1999) times as great as the intended nominal alpha level; the crossing SIBTEST procedure displayed a Type I error rate that was on the order of two times as great the nominal alpha level (Narayanan & Swaminathan, 1996); and the split Mantel-Haenszel procedure displayed a Type I error rate that was in the order of five times as great as the nominal alpha level (Marañón, García, & Costas, 1997). In addition, some of these procedures, such as the logistic regression method, require iterative parameter estimation and thus are computationally demanding. These disadvantages suggest the need for a computationally simple procedure for assessing nonuniform DIF that maintains an acceptable Type I error rate and relatively high power under a variety of conditions.

Breslow and Day (1980) proposed a method for assessing trends in odds ratio heterogeneity that can be applied to the analysis of nonuniform DIF. This procedure has a noniterative solution and may prove to have a power and Type I error rate that is superior to those of earlier proposed methods. This article describes the Breslow-Day procedure and its application to the detection of nonuniform DIF. In addition, because pilot simulations showed that the Breslow-Day procedure displayed strong power under conditions for which the Mantel-Haenszel chi-square did not, and the Mantel-Haenszel chi-square

displayed strong power under conditions for which the Breslow-Day procedure did not, a procedure using the results of both methods in combination to make statistical decisions about DIF is also proposed. The results of a simulation study examining the power and Type I error rates of these procedures are reported and compared with results obtained in earlier research for logistic regression and crossing SIBTEST.

The Breslow-Day and Related Procedures

The Breslow-Day Procedure

Let us estimate ability using the total test score X and denote a particular stratum of test score by k, where $k = 1, \ldots, K$. Then the responses to the studied item of the N_k individuals at the kth stratum can be organized in a 2 x 2 table as shown in Table 1.

The relative performance of the reference and focal groups at the kth stratum can be assessed by considering the ratio of odds of correct response for the members of the two groups at stratum k, ψ_k. This odds ratio can be estimated using

$$\hat{\psi}_k = \frac{a_k d_k}{b_k c_k} \quad .$$

One method to investigate the presence of nonuniform DIF compares the observed values of a_k to those expected under the null hypothesis of uniform DIF. This method is developed as follows. If we treat the marginal values at each stratum as fixed, then at the kth stratum the distribution of a_k follows a noncentral hypergeometric distribution given by

$$P(a_k \mid n_{Rk}, n_{Fk}, n_{1k}, n_{0k}) = \frac{\binom{n_{Rk}}{a_k}\binom{n_{Fk}}{n_{1k}-a_k}\psi_k^{a_k}}{\sum_{u=1}^{U}\binom{n_{Rk}}{u}\binom{n_{Fk}}{n_{1k}-u}\psi_k^{u}} \quad , \tag{1}$$

where ψ_k is the noncentrality parameter (which is equal to the odds ratio at stratum k), and u takes on all possible values of a_k given the configuration of the marginal totals, namely, $(0, n_{1k} - n_{Fk}) \leq u \leq (n_{1k}, n_{Rk})$. When $\psi_k = 1$, Equation 1 reduces to the familiar hypergeometric distribution. If the number of observations at stratum k are relatively large, we can approximate the distribution of a_k by a normal distribution, centered at the expected value of a_k. Let us denote the asymptotic expectation for a_k by A_k, and similarly for the other three cells of Table 1 by B_k, C_k, and D_k. For fixed marginals, once A_k is known, the expected values of the other three cells can be obtained by

$B_k = n_{Rk} - A_k$
$C_k = n_{1k} - A_k$, and
$D_k = n_{Fk} - n_{1k} + A_k$.

Assuming that ψ is constant across all K strata, and using these expectations, the asymptotic odds ratio at stratum k can be expressed as

$$\psi = \frac{A_k D_k}{B_k C_k} = \frac{A_k(n_{Fk} - n_{1k} + A_k)}{(n_{Rk} - A_k)(n_{1k} - A_k)} \quad . \tag{2}$$

Table 1
Responses to the Studied Item by Group Membership

Group	Response 1	0	Total
Reference	a_k	b_k	n_{Rk}
Focal	c_k	d_k	n_{Fk}
Total	n_{1k}	n_{0k}	N_k

Expanding Equation 2 and setting the result equal to zero yields

$$A_k^2(1 - \psi) + A_k(n_{Fk} - n_{1k} + \psi n_{Rk} + \psi n_{1k}) - \psi n_{Rk} n_{1k} = 0 , \quad (3)$$

which is quadratic with respect to A_k. If the asymptotic odds ratio (ψ) is known, then A_k can be obtained by solving the quadratic equation shown in Equation 3 using

$$A_k = \frac{n_{1k} - n_{Fk} - \psi(n_{1k} + n_{Rk}) \pm \sqrt{(n_{Fk} - n_{1k} + \psi n_{1k} + \psi n_{Rk})^2 + 4(1 - \psi)(n_{Rk} n_{1k}\psi)}}{2(1 - \psi)} .$$

Only one root yields possible values of A_k in the sense that A_k, B_k, C_k, and D_k are all nonnegative. Note that A_k represents the expected value of a_k under the assumption of homogeneous odds ratios, which is equivalent to the condition of uniform DIF.

As the odds ratios become more heterogeneous (the nonuniformity of DIF increases), we expect larger deviations between a_k and the expected value under the assumption of homogeneity (A_k). Thus a test for nonuniform DIF can be constructed by considering the deviations between a_k and A_k. This strategy was adopted by Breslow and Day (1980), giving the test statistic

$$BD = \frac{\left[\sum_{k=1}^{K} X_k(a_k - A_k)\right]^2}{\sum_{k=1}^{K} X_k^2 V(a_k) - \frac{\left[\sum_{k=1}^{K} X_k V(a_k)\right]^2}{\sum_{k=1}^{K} V(a_k)}} \quad (4)$$

where X_k is the value of the kth level of the stratifying variable, A_k is the asymptotic expected value of a_k given by the solution of Equation (3), and $V(a_k)$ is the asymptotic variance of a_k given by

$$V(a_k) = \left[\frac{1}{A_k} + \frac{1}{B_k} + \frac{1}{C_k} + \frac{1}{D_k}\right]^{-1} .$$

The statistic BD is the Breslow-Day test for trend in odds ration heterogeneity (Breslow & Day, 1980) and is distributed approximately as chi-square with one

degree of freedom. If the *X*s are interval or ratio in scale, a continuity correction can be used in the numerator before squaring. In the context of DIF detection, *X* represents the level of proficiency, which can be estimated using the total test score.

In computing A_k several options can be used to estimate ψ. The unconditional maximum likelihood estimate can be used, but has the drawbacks of requiring large stratum sizes and being biased, with an asymptotic value equaling the square of the population odds ratio (Breslow & Day, 1980). An alternative strategy that is recommended by Breslow and Day is to use the Mantel-Haenszel estimate of the common odds ratio. Descriptions of the computation and interpretation of the Mantel-Haenszel common odds ratio in the context of DIF detection are provided by Camilli and Shepard (1994) and Dorans and Holland (1993). The Mantel-Haenszel common odds ratio has the advantages of being computationally simple and asymptotically unbiased. This article uses the Mantel-Haenszel estimate of the common odds ratio in computing BD.

It should be noted that Breslow and Day (1980) also proposed a global test of odds ratio heterogeneity, *T*, that is distributed approximately as chi-square with *K*–1 degrees of freedom when the number of observations per stratum is large. The statistic *T* has several known drawbacks. First, because it is a global statistic it cannot assess the specific alternative that there is a systematic increase or decrease in the odds ratios across the ability continuum, as would be the case in nonuniform DIF. Second, if there are relatively few observations per stratum, *T* may not approximate the nominal chi-square distribution, even when the null hypothesis of homogeneity holds. Pilot simulations showed that unless the sample sizes were very large (e.g., > 2000), *T* had very low power, rarely exceeding .20. As a result, the utility of *T* in detecting nonuniform DIF is not pursued in this article.

The Combined Decision Rule Procedure

The Mantel-Haenszel chi-square (MH) is known to be the most powerful test of uniform DIF (Cox, 1988), but has been shown to be relatively ineffective at detecting crossing-nonuniform DIF when the item difficulty was medium (Narayanan & Swaminathan, 1996; Swaminathan & Rogers, 1990). The formula required for computing MH is presented in the context of DIF detection by Camilli and Shepard (1994). Because pilot simulations indicated that the power of BD for detecting crossing-nonuniform DIF tended to be relatively high when the difficulty of the studied item was medium, but decreased substantially as the difficulty of the studied item became more extreme, it was of interest to determine the extent to which a decision rule based on a combination of the individual decisions made according to BD and MH could maintain high power and adequate Type I error rates across all levels of studied item difficulty. This combined decision rule is denoted here by CDR. The CDR accepts the null hypothesis of no DIF if both BD and MH lead to decisions of accepting the null hypothesis, and the CDR rejects the null hypothesis of no DIF if either BD or MH leads to a decision of rejecting the null hypothesis.

The CDR procedure is based on the results of two statistical tests, and as such the significance level used for each test of the CDR procedure requires correction to obtain the intended nominal Type I error rate. I recommend the

use of the Bonferroni correction (Mendenhall, Scheaffer, & Wackerly, 1986), whereby the intended nominal Type I error rate is divided by the number of individual tests conducted (in this case, two) to arrive at the per-test significance level for the CDR procedure. That is, if a nominal Type I error rate is set to .05 ($\alpha = .05$), then the CDR procedure assesses DIF using BD and MH with $\alpha = .025$. The Bonferroni correction is commonly used to adjust the per-test significance level when multiple tests are conducted on the same data (Keppel, 1991) and has also been proposed to adjust the per-test significance level when multiple tests of DIF are conducted on the same item (Penfield, 2001).

Method

A simulation study was conducted to assess the power and Type I error rate of BD, MH, and CDR under a variety of conditions. A secondary purpose of the simulation study was to permit a comparison of the power and Type I error rate of BD and CDR to that observed for crossing SIBTEST and logistic regression as reported in Narayanan and Swaminathan (1996). To this end, many of the properties of the simulation study conducted here follow the methods employed by Narayanan and Swaminathan. Although there are some differences in the simulation methods used here to those of Narayanan and Swaminathan, the methods were viewed as being similar enough to permit a clear comparison of the performance of crossing SIBTEST and logistic regression to the procedures presented here.

Simulation Procedures

An artificial test was constructed of 40 dichotomous items. One of the 40 items, the studied item, was tested for DIF using BD, MH, and CDR. For each item, responses were generated by (a) drawing a random variate from a normal distribution with mean μ and standard deviation of one, (b) determining the probability (P) of correct response on the item according to the three-parameter IRT model (Lord, 1980), (c) drawing a random variate (U) from a uniform distribution on the interval 0 to 1, and (d) assigning a response of 1 if $P \geq U$ and 0 if $P < U$.

The parameters of the nonstudied items were assigned as follows: each c parameter value was set to 0.2, each b parameter was sampled from $N(0, 1)$, and each a parameter was set to $\exp(z)$, where z was sampled from $N(0, 0.1225)$. These parameter distributions are consistent with those used in earlier research and represent realistic distributions of item parameters (Donoghue & Allen, 1993; Zwick, Donoghue, & Grima, 1993). For the studied item, the c parameter was assigned a value of 0.2, and the b and a parameters were fixed as described below.

To investigate the performance of BD, MH, and CDR under a variety of conditions, the following factors were manipulated: DIF effect size, equality of group ability distributions, sample size, the difficulty of the studied item, and the discrimination of the studied item. Each of these factors is discussed below.

Magnitude of nonuniform DIF. The magnitude of nonuniform DIF introduced into the studied item was determined by the area between the item characteristic curves of the reference and focal groups. This area is used as an index of DIF effect size (represented by Δ) and is obtained using the derivations provided by Raju (1988). Five levels of effect size were used: $\Delta = 0.0$, $\Delta = 0.4$, $\Delta = 0.6$,

$\Delta = 0.8$, and $\Delta = 1.0$. The case of $\Delta = 0$ is used to assess the Type I error rates of BD, MH, and CDR. For each effect size, two sets of *a* parameters were used; one set where the *a* parameters were relatively low (Low-*a* condition), and one set where the *a* parameters were relatively high (High-*a* condition). The values of the *a* parameters in the low and high conditions are displayed in Table 2. For $\Delta = 0.4$, 0.6, 0.8, and 1.0 the values of a_F and a_R are identical to those used by Narayanan & Swaminathan (1996). For $\Delta = 0.0$, the values of a_F and a_R were both set to 0.66 under the Low-*a* condition and 1.26 under the High-*a* condition. The value of 0.66 equals the mean of all eight values of a_F and a_R in the Low-*a* condition, and the value of 1.26 equals the mean of all eight values of a_F and a_R in the High-*a* condition.

Equality of group ability distributions. It is often the case that the ability distributions of the reference and focal groups have unequal means. To examine the effect of the equality of group ability distributions on the procedures being studied, two levels of equality were used. In the first level the groups had equal means ($\mu_R = \mu_F = 0$), and in the second level the focal group had a mean that was one standard deviation below the mean of the reference group ($\mu_R = 0$ and $\mu_F = -1$).

Group size. Four levels of group size were investigated in this study: (a) $N_R = 500$ and $N_F = 200$, (b) $N_R = 500$ and $N_F = 500$, (c) $N_R = 1{,}000$ and $N_F = 200$, and (d) $N_R = 1{,}000$ and $N_F = 500$. These levels are representative of sample sizes found in practical testing situations, and are identical to those employed by Narayanan and Swaminathan (1996).

Difficulty of the studied item. Three levels of difficulty of the studied item were investigated: $b = -1.5, 0.0, 1.5$. These levels are identical to those employed by Narayanan and Swaminathan (1996).

Discrimination of the studied item. Two levels of discrimination of the studied item were investigated: a low discrimination level (Low-*a*) and a high discrimination level (High-*a*). The values of the *a* parameters for the focal and reference groups varied depending on the level of effect size, as displayed in Table 2. However, for each level of effect size, there are two levels of *a* parameters corresponding to the Low-*a* and High-*a* conditions.

This design yielded 240 conditions (5 levels of effect size x 2 levels of ability distribution equality x 4 levels of sample size x 3 levels of studied item difficulty x 2 levels of studied item discrimination). Each condition was replicated 1,000 times, and across the 1,000 trials the proportion of trials for which the null hypothesis of no DIF was rejected was recorded for BD, MH, and CDR procedures. These proportions serve as estimates of the power and Type I error rates of each procedure. The nominal Type I error rate used for the tests of DIF using BD and MH in isolation was .05, and the nominal Type I error rate used for BD and MH in the CDR test of DIF was .025 (thus employing the Bonferroni correction). In all conditions, $b_R = b_F$ for all items, including the studied item, and thus all simulated DIF was of the crossing-nonuniform type.

Accommodating Empty Strata

It is occasionally the case that strata contain no data for either the reference group or the focal group, particularly for strata at the extreme ends of the test score continuum. For such strata the calculation of an odds ratio is impossible due to the presence of fractions with denominators with values of zero. A

Table 2
Values of the *a* Parameters for the Studied Item

	Low-a		High-a	
Effect Size (Δ)	a_F	a_R	a_F	a_R
0.0	0.66	0.66	1.26	1.26
0.4	0.72	0.50	2.01	0.90
0.6	0.80	0.46	1.97	0.70
0.8	0.91	0.43	1.79	0.56
1.0	1.03	0.40	1.68	0.47

Note. For the condition in which Δ = 0, the values of *a* equal the mean of the *a* values in other four effect size condition across both the reference and focal groups.

commonly employed strategy to accommodate empty cells is to add the value of 0.5 to each cell of Table 1, thus avoiding zero denominators (Agresti, 1990). This strategy is inadequate in the context of DIF, as it leads to grossly inflated Type I error rates of contingency table methods of DIF detection (such as MH and BD) when the ability distributions of the reference and focal groups differ. As a result, this simulation study employed the strategy of omitting from the analysis the data from any strata for which either the reference or focal group had a zero frequency.

Differences from the Methods of Narayanan and Swaminathan (1996)

Although this study used a simulation procedure that was nearly identical to that of Narayanan and Swaminathan (1996), there were two subtle differences. First, the values of the *a* parameters when Δ = 0 were not reported in Narayanan and Swaminathan (1996), and thus were probably not identical to those used here. However, because the values of the *a* parameters used here when Δ = 0 were equal to the average value of those used in the other levels of effect size, it seems likely that similar values would have been used by Narayanan and Swaminathan. Second, Narayanan and Swaminathan investigated the effect of matching criterion contamination, and thus introduced nonuniform DIF into varying numbers of items on the test. For realistic levels of contamination (10%-20% of the items displaying DIF), this effect was shown to have little effect on the power and Type I error rate of logistic regression, crossing SIBTEST, and MH (Narayanan & Swaminathan), as well as on other DIF detection procedures (Penfield, 2001). As a consequence, the effect of matching criterion contamination was not expected to have a substantial effect on the performance of CDR, and was not studied in this simulation.

Results

Type I Error Rates

Table 3 presents the Type I error rates for BD, MH, and CDR as a function of group size, item type, and equality of the reference and focal group ability distributions. Inspection of the Type I error rates indicates that all procedures maintained Type I error rates that were consistently at or below the nominal level of .05 when group ability distributions were equal, and consistently near the nominal level of .05 when group ability distributions were unequal. In

Table 3
Type I Error Rate as a Function of Sample Size and Item Type

Factor	$\mu_R - \mu_F = 0$ BD	MH	CDR	$\mu_R - \mu_F = 1$ BD	MH	CDR
Sample Size						
$N_R = 500$, $N_F = 200$	.01	.05	.04	.02	.06	.04
$N_R = 500$, $N_F = 500$	.01	.05	.03	.02	.07	.05
$N_R = 1{,}000$, $N_F = 200$	.02	.06	.04	.02	.07	.04
$N_R = 1{,}000$, $N_F = 500$	.02	.06	.03	.02	.08	.05
Item Type						
Low-*b*, High-*a*	.00	.05	.03	.00	.09	.04
Med-*b*, Low-*a*	.02	.06	.04	.02	.06	.04
Med-*b*, High-*a*	.00	.05	.03	.01	.06	.03
High-*b*, Low-*a*	.04	.06	.04	.05	.08	.07

Note. BD, MH, and CDR correspond to the Breslow-Day test, the Mantel-Haenszel chi-square, and the combined decision rule respectively.

general, MH and CDR displayed Type I error rates that were higher than that of BD, particularly in the conditions for which the discrimination of the studied item was high (High-*a*). Of most importance to this study is the observation that the Type I error rate of CDR remained at or below the nominal level of .05 for all conditions except one (High-*b*, Low-*a*, $\mu_R - \mu_F = 1$), for which the Type I error rate equaled .07.

Power

Table 4 displays the power of BD, MH, and CDR as a function of group size, effect size, item type, and difference between the ability distributions of the reference and focal group. These results indicate three general trends in the power rates. First, the power of BD was relatively low when the studied item difficulty was extreme (Low-*b* and High-*b*) and relatively high when the studied item difficulty was moderate (Med-*b*). Second, the power of MH showed an inverse relationship to that of BD, being very low when the studied item difficulty was moderate, and relatively high when the studied item difficulty was extreme. This result confirms the findings of pilot simulations and suggests that a combined use of BD and MH would provide high power across all levels of studied item difficulty. The third result of interest is that the power of CDR was consistently high across all levels of studied item difficulty and discrimination. The power of CDR tended to be higher for High-*a* than Low-*a* conditions, and higher for conditions with equal group ability distributions than unequal group ability distributions. In addition, the power for all three procedures was more dependent on the presence of a small group size than on the combined number of reference and focal group members; that is, the power was higher for $N_R = N_F = 500$ (combined group size of 1,000) than for $N_R = 1{,}000$ and $N_F = 200$ (combined group size of 1,200).

Table 4
Power as a Function of Sample Size, Effect Size, and Item Type

Factor	$\mu_r - \mu_F = 0$ BD	MH	CDR	$\mu_r - \mu_F = 1$ BD	MH	CDR
Sample Size						
$N_R = 500$, $N_F = 200$	.30	.39	.50	.20	.33	.36
$N_R = 500$, $N_F = 500$	.45	.44	.71	.37	.45	.54
$N_R = 1{,}000$, $N_F = 200$	.40	.41	.58	.30	.36	.45
$N_R = 1{,}000$, $N_F = 500$	.57	.47	.80	.50	.49	.66
Effect Size						
$\Delta = 0.4$	.17	.32	.37	.10	.21	.20
$\Delta = 0.6$	.36	.41	.59	.27	.37	.43
$\Delta = 0.8$	.54	.47	.76	.45	.48	.62
$\Delta = 1.0$	.65	.51	.87	.56	.57	.76
Item Type						
Low-*b*, High-*a*	.43	.96	.95	.46	.81	.78
Med-*b*, Low-*a*	.52	.06	.45	.37	.10	.32
Med-*b*, High-*a*	.72	.07	.65	.51	.23	.50
High-*b*, Low-*a*	.05	.62	.55	.04	.49	.42

Note. BD, MH, and CDR correspond to the Breslow-Day test, the Mantel-Haenszel chi-square, and the combined decision rule respectively.

Comparing CDR to Logistic Regression and Crossing SIBTEST

The results of the simulation study indicated that although BD and MH were effective in detecting nonuniform DIF under specific conditions, CDR was consistently effective across all conditions. Because CDR appears to hold the greatest potential for improving the available methodology for DIF detection, it was of interest to compare the performance of CDR with other popular methods for detecting nonuniform DIF. Table 5 presents a comparison of the Type I error rate and power of CDR with that obtained for logistic regression and crossing SIBTEST in earlier research (Narayanan & Swaminathan, 1996). The observed Type I rate of CDR was .04 across all conditions, substantially lower than the Type I error rates of approximately .09 observed for logistic regression and crossing SIBTEST across a nearly identical set of conditions. As a result, CDR appears to have a clear advantage in terms of Type I error rate.

With respect to power, the values obtained for CDR displayed in Table 5 are slightly lower than those obtained for logistic regression and crossing SIBTEST. The average difference in power between CDR and logistic regression was .05, and between CDR and crossing SIBTEST was .10. The largest discrepancies in power between CDR and the other two procedures occurred when group sizes were small (N_R = 500 and N_F = 200) and the studied item difficulty and discrimination were medium and high, respectively (Med-*b*, High-*a*). When the group sizes were larger (N_R = 1,000 and N_F = 500), CDR displayed a power that was only .02 below that of logistic regression and .06 below that of crossing SIBTEST.

Table 5
Comparing the Type I Error Rate and Power of CDR to Logistic Regression and Crossing SIBTEST

Factor	Type I Error Rate			Power		
	CDR	LR	SIB	CDR	LR	SIB
Sample Size						
$N_R = 500$, $N_F = 200$	.04	.08	.08	.43	.52	.58
$N_R = 500$, $N_F = 500$	.04	.08	.09	.63	.70	.72
$N_R = 1{,}000$, $N_F = 200$	.04	.09	.08	.52	.57	.62
$N_R = 1{,}000$, $N_F = 500$	.04	.09	.09	.73	.75	.79
Item Type						
Low-*b*, High-*a*	.04	.10	.11	.87	.90	.88
Med-*b*, Low-*a*	.04	.08	.08	.39	.44	.47
Med-*b*, High-*a*	.03	.09	.09	.58	.70	.77
High-*b*, Low-*a*	.05	.06	.07	.49	.48	.59

Note. Reported values for sample size were obtained by collapsing across all conditions other than sample size, and reported values for item type were obtained by collapsing across all conditions other than item type. Values for logistic regression (LR) and crossing SIBTEST (SIB) were obtained from Narayanan and Swaminathan (1996), in which 100 trials were run for each condition.

Discussion

This study investigated the power and Type I error rate of the Breslow-Day test of trend in odds ratio heterogeneity (BD) and a combined decision rule (CDR) that is based on the outcomes of both BD and the Mantel-Haenszel chi-square (MH). A simulation study led to two general results regarding the power and Type I error rate of BD and CDR. First, BD used in isolation maintained a subnominal Type I error rate across all conditions and a high power when the studied item difficulty was moderate. Second, CDR maintained a Type I error rate at or below the nominal level across nearly all conditions and also displayed consistently high power across all levels of studied item difficulty. A comparison of the power and Type I error rate of CDR with those obtained for logistic regression and crossing SIBTEST in earlier research (Narayanan & Swaminathan, 1996) indicated that CDR displayed a Type I error rate that was consistently more than 50% lower than that of the other two procedures (.04 vs. approximately .09) and that CDR displayed a power that was slightly lower than that of the other procedures.

The results of this study clearly indicate that CDR has a Type I error rate that is superior to that of logistic regression and crossing SIBTEST. However, the inflated Type I error rates of logistic regression and crossing SIBTEST hamper a comparison of their power with that of CDR and thus prevent a clear conclusion concerning their relative effectiveness in practice. Narayanan and Swaminathan (1996) provided an estimate of the significance level required for logistic regression and crossing SIBTEST to obtain an observed Type I error rate of .05. For both procedures the corrected significance level appears to be approximately .025. It is currently unknown what the power of logistic regres-

sion and crossing SIBTEST would be if this corrected significance level were used, but it is expected to be lower than that reported by Narayanan and Swaminathan. My simulations showed that a change in the significance level from .05 to .025 led to a decrease in power of BD and MH by approximately .10 (although this value varied greatly depending on the condition, sometimes being as little as .01 and other times being as great as .20). If the power of logistic regression and crossing SIBTEST are affected similarly to that of BD and MH, then the power of logistic regression and crossing SIBTEST using the corrected significance level would be expected to be lower than that of CDR. Further research is required to understand better how the power of logistic regression and crossing SIBTEST using a corrected significance level compares with that of CDR.

This study demonstrated the good Type I error rate and relatively strong power exhibited by the CDR procedure for detecting crossing-nonuniform DIF. This result, in combination with the fact that MH is known to be the most effective method for detecting uniform DIF (Camilli & Shepard, 1994; Cox, 1988), suggests that CDR may be the single most effective method for simultaneously detecting uniform and nonuniform DIF in dichotomous items. Because CDR has the added advantage of being computationally simple relative to other procedures that require an iterative parameter estimation algorithm, such as logistic regression and IRT methods, CDR appears to be an attractive alternative to other procedures currently in use.

The use of CDR in DIF detection analyses may improve the ability of test developers to identify items that contain bias. Because currently employed methods of DIF detection are poor at simultaneously detecting uniform and nonuniform DIF, CDR offers test developers a means to assess each item for a large variety of forms of DIF and thus a highly powerful method of assessing each item for bias. The use of CDR to improve the detection of biased test items may ultimately lead to an improvement in the validity of test scores.

References

Agresti, A. (1990). *Categorical data analysis*. Toronto, ON: Wiley.

Breslow, N.E., & Day, N.E. (1980). *Statistical methods in cancer research: Volume 1—The analysis of case-control studies*. Lyon: International Agency for Research on Cancer.

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.

Cox, D.R. (1988). *Analysis of binary data* (2nd ed.). London: Methuen.

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.

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

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

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*, 313-334.

Hanson, B.A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23*, 244-253.

Hills, J.R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, 8*(4), 5-11.

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

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

Li, H., & Stout, W.F. (1993, April). *A new procedure for detection of crossing DIF/bias.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Marañón, P.P., García, M.I.B., & Costas, C.S.L. (1997). Identification of nonuniform differential item functioning: A comparison of Mantel-Haenszel and item response theory analysis procedures. *Educational and Psychological Measurement, 57,* 559-568.

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.

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

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

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

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.

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

Shealy, R.T., & Stout, W.F. (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,* 197-239.

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

Whitmore, M.L., & Schumacker, R.E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement, 59,* 910-927.

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