Research Report

Three Statistical Testing Procedures in Logistic Regression: Their Performance in Differential Item Functioning (DIF) Investigation

Insu Paek

December 2009

ETS RR-09-35

Listening. Learning. Leading.®

Three Statistical Testing Procedures in Logistic Regression: Their Performance in Differential Item Functioning (DIF) Investigation

Insu Paek

ETS, Princeton, New Jersey

December 2009

Abstract

Three statistical testing procedures well-known in the maximum likelihood approach are the Wald, likelihood ratio (LR), and score tests. Although well-known, the application of these three testing procedures in the logistic regression method to investigate differential item function (DIF) has not been rigorously made yet. Employing a variety of simulation conditions, this research (a) assessed the three tests' performance for DIF detection and (b) compared DIF detection in different DIF testing modes (targeted vs. general DIF testing). Simulation results showed small differences between the three tests and different testing modes. However, targeted DIF testing consistently performed better than general DIF testing; the three tests differed more in performance in general DIF testing and nonuniform DIF conditions than in targeted DIF testing and uniform DIF conditions; and the LR and score tests consistently performed better than the Wald test.

Key words: DIF, Mantel-Haenszel statistic, logistic regression, Wald test, Likelihoood ratio test, score test

Table of Contents

Page

List of Tables

List of Figures

Page

Differential item functioning (DIF) has been a popular research topic in the measurement and testing field. Many DIF methods have been investigated and developed, including the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988; Mantel & Haenszel, 1959), the item response theory (IRT) parameter chi-square test (Lord, 1977, 1980), the IRT likelihood ratio approach (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993), and the simultaneous item bias test (SIBTEST; Shealy & Stout, 1993). (For reviews of various DIF methods and details on their classifications, see Camilli and Shepard, 1994; Dorans and Potenza, 1994; Holland and Wainer, 1993; and Millsap and Everson, 1993.) Another well-known DIF method for dichotomously scored items is the logistic regression DIF procedure introduced by Swaminathan and Rogers (1990), who showed that logistic regression can be used to detect uniform and nonuniform DIF.[1]

The following three models (M0, M1, and M2) are of major interest with respect to the logistic regression DIF procedure. Let *Y* be a response to a given item (*0* for incorrect, *1* for correct); the expectation of *Y* (the probability of a correct response to the item) is $E(Y) = \pi = \frac{e^{\eta}}{1+e^{\eta}}$ where η has the following forms in which *T* is the test score (total number correct), *G* is a group indicator variable, and (*TG*) is a product of *T* and *G*.

$$\text{M0: } \eta = \beta_0 + \beta_1 T \tag{1}$$

$$\text{M1: } \eta = \beta_0 + \beta_1 T + \beta_2 G \tag{2}$$

$$\text{M2: } \eta = \beta_0 + \beta_1 T + \beta_2 G + \beta_3 (TG) \tag{3}$$

The comparisons of interest in DIF detection are M0 versus M1 (a test of uniform DIF), M1 versus M2 (a test of nonuniform DIF),[2] and M0 versus M2 (a test of any type of DIF, uniform or nonuniform). Swaminathan and Rogers (1990) introduced a chi-square test to statistically test DIF by comparing M0 and M2. Although unnamed in their report, this comparison is the Wald test. At least two other statistical tests are also readily available for the logistic regression procedure: the LR test and the score test (Lagrange multiplier test). The Wald, LR, and score tests are asymptotically equivalent (Cox & Hinkley, 1974). Which of the three tests is preferable depends on the situation. However, there has been little information or consensus regarding their comparative performance in detecting DIF by the logistic regression procedure.

The LR test was mainly introduced and used for IRT DIF analysis (e.g., Thissen et al., 1993), and as far as the author knows, only Glas (1998) has used the score test for that purpose. Some evidence suggests that the Wald test has weak statistical power. Hauck and Donner (1977) showed that in single-parameter testing for the binomial logit model, the Wald statistic decreases and its power becomes weaker when the true distance between the null hypothesis and the alternative hypothesis becomes large; ultimately its power diminishes to the significance level. Comparing the Wald and LR tests, Hauck and Donner recommended the LR test. Vaeth (1985) studied the aberrant behavior of the Wald test when it is used for hypothesis testing in exponential families. He concluded that the Wald test requires caution when applied to logistic regression with many predictors. Fears, Benichou, and Gail (1996) showed that the Wald test power was weaker than the usual *F* test in an application of random-effects analysis of variance. Pawitan (2000) explained the Wald test's lack of power in terms of maximum likelihood perspective . The equivalence of the score and MH tests (Day & Byar, 1979) is also relevant to this paper. In sum, it is useful to compare the statistical performances of the Wald, LR, and score tests with regard to logistic regression DIF application.

Swaminathan and Rogers' (1990) general DIF test compares M0 and M2 but this general DIF testmay not be as statistically powerful as more-targeted tests for uniform DIF (M0 vs. M1) or nonuniform DIF (M1 vs. M2) because one degree of freedom is lost by modeling β_3 in M2. Sometimes practitioners and researchers use the logistic regression DIF procedure to investigate only uniform DIF (e.g., Monahan, McHorney, Stump, & Perkins, 2007), even though the procedure is also designed fordetecting nonuniform DIF. Therefore, it would be useful to know the extent to which targeted and general DIF testing differ in their power to detect DIF.

Through systematically varied simulations, this study investigated the DIF-detection capabilities of the Wald, LR, and score tests. It also compared different DIF testing modes:: generaltesting (M0 vs. M2) and targeted DIF testing (uniform DIF [M0 vs. M1] and nonuniform DIF [M1 vs M2]).). The simulation varied sample size and DIF magnitudes for item response data generation. In sum, the research focused on differences in DIF detection (a) between the Wald, LR, and score tests and (b) between different DIF testing modes(targeted vs. general). The study also investigated which sample sizes ensure adequate statistical power to detect items with medium or large DIF for the three tests and the different testing modes, given the simulation conditions in this study.

Three Statistical Tests for Differential Item Functioning

The logistic regression is one of the generalized linear models in which statistical testing is based on maximum likelihood (ML) estimation. The Wald, LR, and score tests are three common ways of testing hypotheses for model parameters or model comparisons in a generalized linear model. (For details on these tests, see Cox & Hinkley, 1974; Dobson, 2002; and Harrell, 2001.) With regard to these three tests, let $\boldsymbol{\beta}$ be a k x 1 vector for the parameters of some general model that consists of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, where $\boldsymbol{\beta}_1$ is a nuisance parameter vector with p number of elements and $\boldsymbol{\beta}_2$ is a vector for parameters with q number of elements for testing the hypothesis H$_o$: $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^*$, where $\boldsymbol{\beta}_2^*$ is a vector with hypothesized fixed constants.

LR Test

Let $L(\boldsymbol{\beta})$ be the likelihood of a general model and $L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2^*)$ be the likelihood of a nested model under the general model with the restriction of $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^*$. The LR statistic is

$$LR = -2\ln\left[\frac{L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2^*)}{L(\boldsymbol{\beta})}\right], \tag{4}$$

where ln represents the natural log, and the likelihood functions, $L(\boldsymbol{\beta})$ and $L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2^*)$, are evaluated at their ML estimates. The LR statistic in Equation 4 follows an asymptotic chi-square distribution with degrees of freedom (*df*) of $k - p = q$. The LR test requires estimation of both a general model and a nested model.

Wald Test

Let $\mathbf{I}(\boldsymbol{\beta})$ be a general model's Fisher information, defined as

$$\mathbf{I}(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right]. \tag{5}$$

Let $\mathbf{I}^{-1}(\boldsymbol{\beta})$, the variance-covariance matrix of the maximum likelihood (ML) estimator for the general model, be partitioned as follows:

$$\mathbf{I}^{-1}(\boldsymbol{\beta}) = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{11}} & \boldsymbol{\Sigma}_{\mathbf{12}} \\ \boldsymbol{\Sigma}_{\mathbf{21}} & \boldsymbol{\Sigma}_{\mathbf{22}} \end{pmatrix}, \tag{6}$$

where $\boldsymbol{\Sigma}_{\mathbf{11}}$ and $\boldsymbol{\Sigma}_{\mathbf{22}}$ are the variance-covariance matrices for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, respectively, and $\boldsymbol{\Sigma}_{\mathbf{12}}$ (or $\boldsymbol{\Sigma}_{\mathbf{21}}$) is the covariance matrix for $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. The Wald statistic then is

$$W = (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*)' \boldsymbol{\Sigma}_{\mathbf{22}}^{-1} (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*), \tag{7}$$

where $\hat{\boldsymbol{\beta}}_2$ is the ML estimate and $\boldsymbol{\Sigma}_{\mathbf{22}}^{-1}$ is evaluated using its ML estimates. The Wald statistic has an asymptotic chi-square distribution with *df* = *q*. The Wald test is simpler than the LR test in that it does not require nested or reduced model estimation. Only the ML estimates and their variance-covariance matrix from the general model should be fully estimated by ML. Note that $\boldsymbol{\beta}_2$ and $\boldsymbol{\Sigma}_{\mathbf{22}}$ are estimated and are adjusted for all other parameters and variance-covariances in the general model. For uniform DIF detection, only M1 should be estimated. For nonuniform DIF detection and general DIF testing, only M2 should be estimated. Swaminathan and Rogers (1990) used this Wald statistic to assess DIF. This procedure is essentially a chi-square test with *df* = 2, a comparison of M0 (reduced or nested model) and M2 (general model). Swaminathan and Rogers' chi-square test formulation is based on a general linear hypothesis test, H$_o$: $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, where **C** is a contrast matrix to set up the null hypothesis H$_o$: $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^* = \mathbf{0}$. For H$_o$: $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^* = \mathbf{0}$, Equation 7 can be re-expressed using the general linear contrast matrix **C** as

$$W = (\mathbf{C}\boldsymbol{\beta})'(\mathbf{C}\mathbf{I}^{-1}(\boldsymbol{\beta})\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta}). \tag{8}$$

In terms of notation, $\boldsymbol{\beta}$ and $\mathbf{I}^{-1}(\boldsymbol{\beta})$ are equivalent to the $\boldsymbol{\tau}$ and $\boldsymbol{\Sigma}$ used by Swaminathan and Rogers (**C** remains the same).

Score Test

The score test measures the difference of the slope of the log-likelihood, ln(*L*), from zero when ln(*L*) is evaluated using the values of the null hypothesis H$_o$: $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^*$. Let **U** be the first partial derivative of a general model's ln(*L*). This derivative is called a score vector. As before, also let $\mathbf{I}^{-1}$ be the general model's variance-covariance matrix. The score statistic is then

$$S = \mathbf{U}'(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2^*)\mathbf{I}^{-1}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2^*)\mathbf{U}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2^*), \qquad (9)$$

where $\mathbf{U}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2^*)$ and $\mathbf{I}^{-1}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2^*)$ are a score vector and a variance-covariance matrix evaluated at the ML estimates of $\boldsymbol{\beta}_1$ with the restriction of H$_o$: $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^*$. The score statistic follows an asymptotic chi-square distribution with *df* = *q*. The score test requires only ML estimates of the nested model. That is, Equation 9 is evaluated using the nested-model ML estimates and does not need full ML estimates of parameters in the general model. The score test may be more challenging for practitioners than the Wald and LR tests because not all popular statistical packages provide the information needed to compute the score statistic. Basically, the score test requires a first and a second derivative of the log-likelihood function with respect to general model parameters. In matrix form the score function, the first derivative of ln(*L*), can be shown as

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}), \qquad (10)$$

where **X** is a design matrix (including **1** for the intercept), **Y** is a response vector, and $\hat{\mathbf{Y}}$ is the predicted response vector. In matrix form the Fisher information is

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{D}\mathbf{X}, \qquad (11)$$

where **D** is a diagonal matrix of weights defined as

$$\mathbf{D} = diag[\pi_n(1 - \pi_n)], \qquad (12)$$

where π_n (n = 1, 2, 3, …, N) is the predicted probability of the *n*th person's correct response to a given item. McCullagh and Nelder (1989) gave the detailed derivations of the score function and the Fisher information. Because of the simple matrix forms in Equations 10 to 12, the score statistic can easily be calculated in the logistic regression DIF procedure. Pregibon (1982) showed an alternative method of calculating the score statistic for logistic regression. This method employs the generalized Pearson chi-square residuals from the nested model's final estimates and from the general model's one-step iteration (starting from $\hat{\boldsymbol{\beta}}_1$) results. In the logistic DIF procedure, the steps and formulas presented above provide an equally simple way of calculating the score statistic.

All computations and simulations were conducted using the statistical language R (The R Project for Statistical Computing, 2002).

Equivalence of Score Test and Mantel-Haenszel Chi-Square Statistic

Swaminathan and Rogers (1990) stated that testing the null hypothesis $\beta_2 = 0$ in M1 for uniform DIF testing (comparison of M0 and M1) is equivalent to testing the null hypothesis that the common odds ratio equals 1 in the MH procedure defined by Holland and Thayer (1988). Although targeted DIF (here, uniform DIF) can be tested by the Wald test, it is not equivalent to the MH chi-square test statistic. In statistical literature, a derivation exists in which the score-test statistic in logistic regression is equivalent to the MH chi-square test statistic (Day & Byar, 1979). In the following logistic-regression form, Day and Byar modeled a test of independence in the J x 2 x 2 tables (m = 1, 2, 3, …, J [number of strata] and in the mth 2 x 2 table, where the column represents correct and incorrect responses and the row represents the focal and reference groups) as

$$\text{logit}(\pi_m) = \beta_{0m} + \beta_2 G \quad . \tag{13}$$

In doing so, they showed that the score statistic for testing β_2 in Equation 13 can be expressed as

$$\text{score statistic (with one } df\text{)} = \frac{\left[\sum_m \left(R_{rm} - \frac{N_{rm}R_{rm}}{N_{tm}}\right)\right]^2}{\sum_m \left(\frac{N_{rm}N_{fm}R_{tm}W_{tm}}{N_{tm}^3}\right)}, \tag{14}$$

where R_{rm} is the frequency of correct responses in the reference group, N_{rm} and N_{fm} are the row margins (for the reference and focal groups), R_{tm} and W_{tm} are the column margins (for correct and incorrect responses), and N_{tm} is the total frequency. This score statistic is exactly the same as Cochran's (1954) conditional independence test in the J x 2 x 2 tables. Compared to the MH chi-square statistic (Mantel & Haenszel, 1959) formulas shown by Dorans and Holland (1993, p. 40, Equations 5 and 6), the numerator lacks the continuity correction of -.5, and the denominator (the variance of R_{rm}) lacks finite sampling corrections—that is, it has N_{tm}^3 rather than $N_{tm}^2(N_{tm}-1)$. But Day and Byar also mentioned that if the likelihood function used in their derivation is conditioned on the marginal totals of the separate 2 x 2 tables, a similar derivation removes the difference in the denominator between the score statistic and the MH chi-square test. Therefore, for uniform DIF testing (M0 vs. M1), the MH procedure has a much closer relationship with the score test than (a) the LR test or (b) the Wald test used by Swaminathan and Rogers. Of the three tests, the score test may be expected to perform best for uniform DIF because of its equivalence to the MH chi-square test, which is the uniformly most powerful unbiased test of constant odds ratio (Holland & Thayer, 1988).

Simulation Method and Design

The item response function (IRF) used for simulation was the three-parameter logistic (3PL) item response model (Birnbaum, 1968), which has the form

$$P(Y_i = 1|\theta) = g_i + (1-g_i)\frac{\exp[Da_i(\theta-b_i)]}{1+\exp[Da_i(\theta-b_i)]}, \tag{15}$$

where $P(Y_i = 1|\theta)$ is the probability of a correct response for the ith item (i = 1, 2, 3, …, I), D is 1.7, g_i is the (pseudo) guessing parameter, a_i is the discrimination, and b_i is the item difficulty. Two sets of simulation data were generated: uniform DIF and nonuniform DIF. The uniform DIF data were generated by the difference in item difficulty in the 3PL IRF between the focal and reference groups ($b_F - b_R$), with all other parameters remaining the same in both groups. The nonuniform DIF data were generated by making the discrimination parameter of the focal group differ from that of the reference group ($a_F \neq a_R$), with all other item parameters remaining the same in both groups.[3] Each simulated test had 41 items. Item-difficulty parameters ranged from -2 to 2, in increments of 0.1. Item-discrimination parameters were randomly drawn from *normal* (1, 0.3). Guessing parameters were randomly drawn from a uniform distribution on [0, 0.35].

When there was no DIF, the studied item had $b_F = b_R = -.5$, $a_F = a_R = 1$, and $g_F = g_R = 0.2$. The discrimination parameters ranged from 0.50 to 1.57, with a mean of 0.98 and an SD of 0.28. The guessing parameters ranged from 0.01 to 0.35, with a mean of 0.19 and an SD of 0.10. To model a typically observed group-ability difference in DIF analysis, person-ability θs were drawn from *normal* (-.5, 1) for the focal group and from *normal* (0, 1) for the reference group. For uniform DIF, $b_F - b_R$ was systematically manipulated from 0 to 1 in increments of 0.05, producing 21 conditions. For nonuniform DIF, a_F/a_R varied from 0.5 to 1.5 in increments of 0.05, producing 21 conditions.[4] The sample sizes of the reference and focal groups were the same. The sample size per group ranged from 100 to 500 in increments of 100.

Zwick (1990) showed that the same 3PL IRF hypothesis (no DIF) is not necessarily equivalent to the null hypothesis of no DIF tested in the observed score-matching DIF methods, such as the MH procedure with total test-score matching. In this study logistic regression procedure for DIF, used the observed total test score as a matching variable. The negative impact of using the observed total score as a matching variable, however, may not be serious in this studybecause the number of items (41) was relatively large under the simulated group difference of 0.5 (half an SD). Spray and Miller (1992) reported that using observed score as a matching variable in the MH method did not lead to practically serious impact on their DIF investigation when the test had a relatively large number of items (40 items in their study).

For uniform and nonuniform DIF data, targeted DIF testing (M0 vs. M1) and general DIF testing (M0 vs. M2) were conducted using the Wald, LR, and score tests. The number of data simulation conditions was 210 (2 x 21 x 5): each data set (uniform and nonuniform) had 21 levels and 5 sample sizes (100, 200, 300, 400, and 500). For each of the 210 conditions, six statistical tests (Wald, LR, and score tests, each in a targeted and a general DIF testing mode) were applied with 1,000 replications. The nominal alpha level of 0.05 was used for statistical significance.

Results

For fair comparisons of different tests' detection rates or statistical power, the tests' false positive rate (i.e.,Type I errorrates) should be at least approximately the same, because higher levels of Type I errors could mean that a test's statistical power is rather liberal, showing higher statistical power. The Type I error rates were examined for the Wald, LR, and score tests using the results fromthe $b_{\mathrm{F}} - b_{\mathrm{R}} = 0$ and $a_{\mathrm{F}} = a_{\mathrm{R}} = 1$ conditions. On average, the three tests' Type I error occurred in 5–6% of the replications across different testing modes, DIF types, and sample sizes. In any given condition, the smallest observed difference in Type I error rate was 0; the largest observed difference was a 1.2% difference between the LR and Wald tests with nonuniform DIF, a general DIF testing approach (M0 vs. M2), and $N = 300$. Therefore, it was concluded that the comparisons adequately represented the three tests' DIF detection rates. Table 1 summarizes the Type I error rates of the Wald, LR, and score tests. (Anyone desiring details of all results under all conditions may request them from the author.)

Table 1

Summary of Type I Error Rates for the Wald, Likelihood Ratio (LR), and Score Tests

	Test		
Summary statistics	Wald	LR	Score
Average	0.055	0.059	0.058
SD	0.011	0.011	0.011
Minimum	0.038	0.039	0.039
Maximum	0.073	0.077	0.074

Note. All values were obtained by calculating averages across different sample sizes (5 levels), DIF magnitudes (21 levels), DIF types (uniform and nonuniform), and DIF testing modes (targeted and general). Each of 420 conditions (210 simulation data conditions × 2 DIF testing modes) had 1,000 replications.

Figures 1 and 2 depict the general rejection-rate results. Although all x-axis values in the figures are discrete (21 points), they were expressed with lines for simple graphical depiction.

For uniform DIF (Figure 1), the DIF detection rate increases as the sample size increases across different test types (Wald, LR, and score) and different DIF testing modes (targeted or general DIF). As the difference in item difficulty ($b_{\mathrm{F}} - b_{\mathrm{R}}$) increases, so does the detection rate, nearing a 100% rejection rate for all tests and sample sizes when $b_{\mathrm{F}} - b_{\mathrm{R}} = 1$. In general for

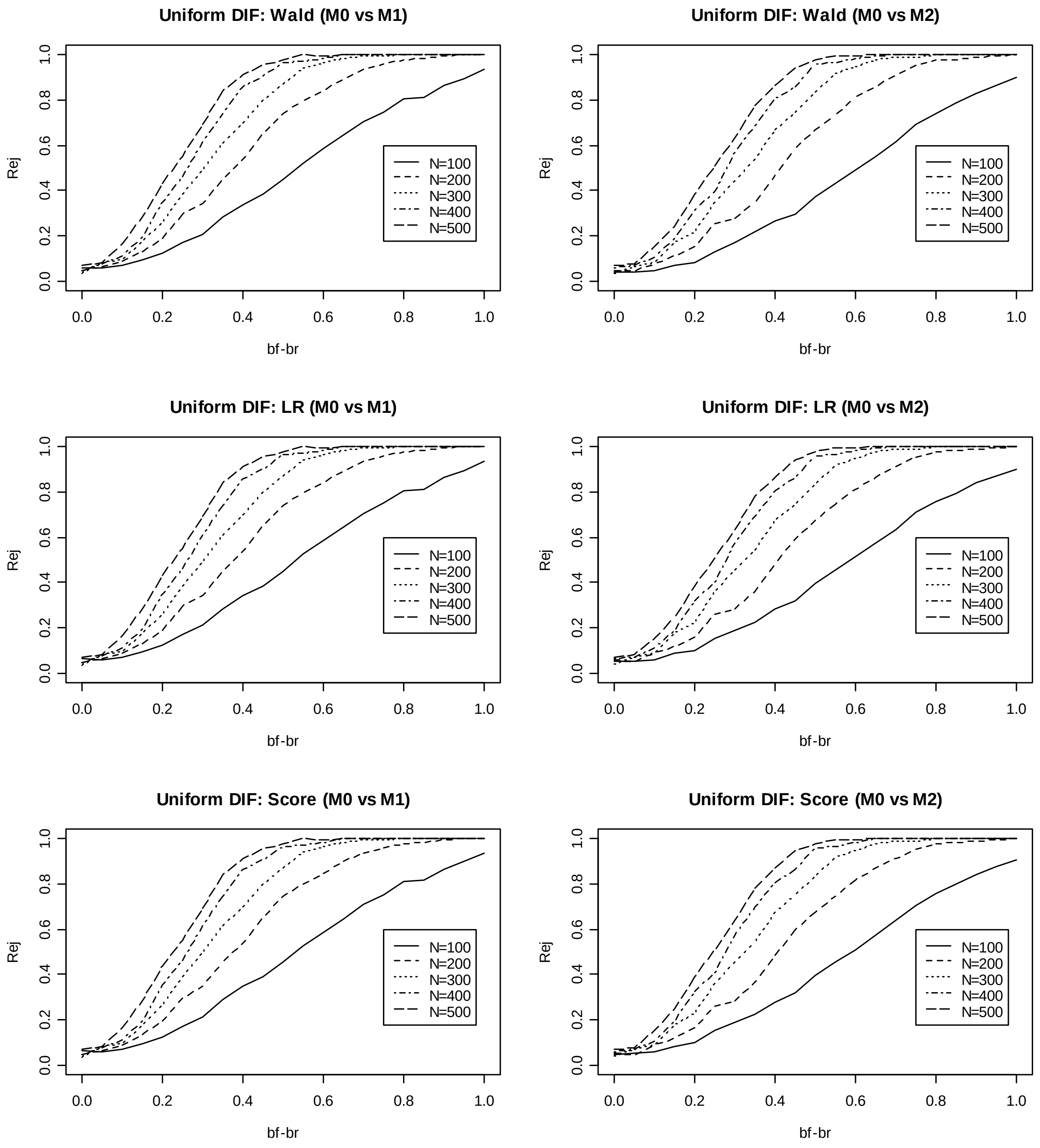


***Figure 1*. Overall rejection rates for uniform differential item functioning (DIF).**

Note. Wald = the Wald test, LR = the likelihood-ratio test, Score = the score test, bf – br = b_F – b_R, and N = sample size per group.

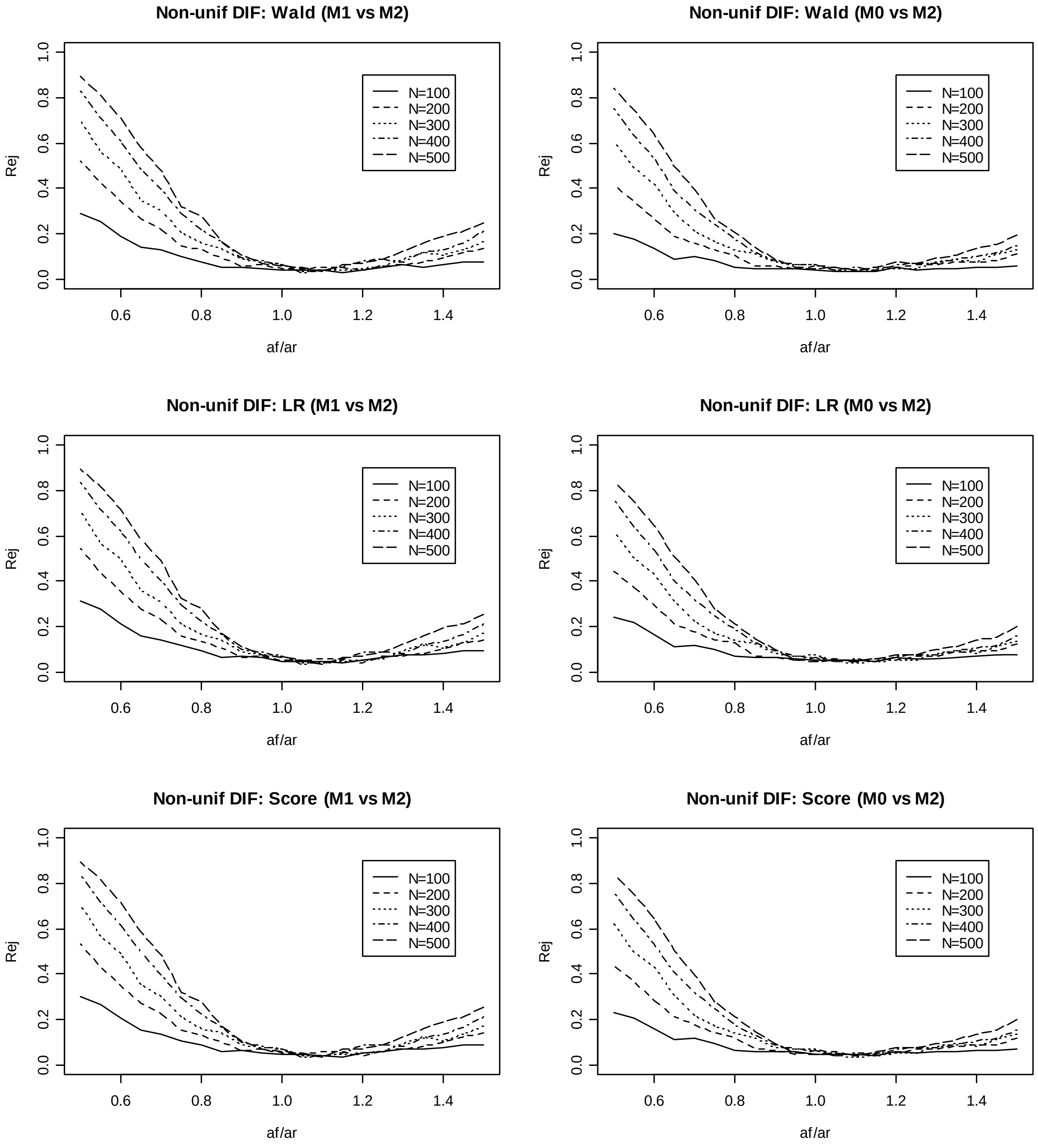


***Figure 2*. Overall rejection rates for nonuniform differential item functioning (DIF).**

Note. Wald = the Wald test, LR = the likelihood-ratio test, Score = the score test, af/ar = a_F/a_R, and N = sample size per group.

nonuniform DIF (Figure 2), increasing the discrimination difference ($a_F/a_R > 1$ or $a_F/a_R < 1$) from no DIF ($a_F = a_R = 1$) brought an increase in rejection rate, but the shape of the rejection-rate pattern was far from a symmetrical U. The DIF detection rate was much higher when $a_F/a_R < 1$ (the focal-group IRF was flatter than the reference-group IRF) than when $a_F/a_R >1$ (the focal-group IRF was steeper than the reference-group IRF). The same discrimination difference did not shape the DIF detection rate symmetrically, indicating that the same differences between a_F and a_R do not cause the same IRF differences.[5] Figure 3 illustrates the same slope differences' nonsymmetrical impact on IRF differences. In the figure, $a_F/a_R = 0.5$ results in a larger absolute difference in area between IRFs than $a_F/a_R = 1.5$.

Comparisons of Wald, Likelihood Ratio, and Score Tests

Using the Wald test's rejection rates as baseline, I compared the rejection rates of the Wald, LR, and score tests. Figures 4–7 show the difference plots for uniform and nonuniform DIF.

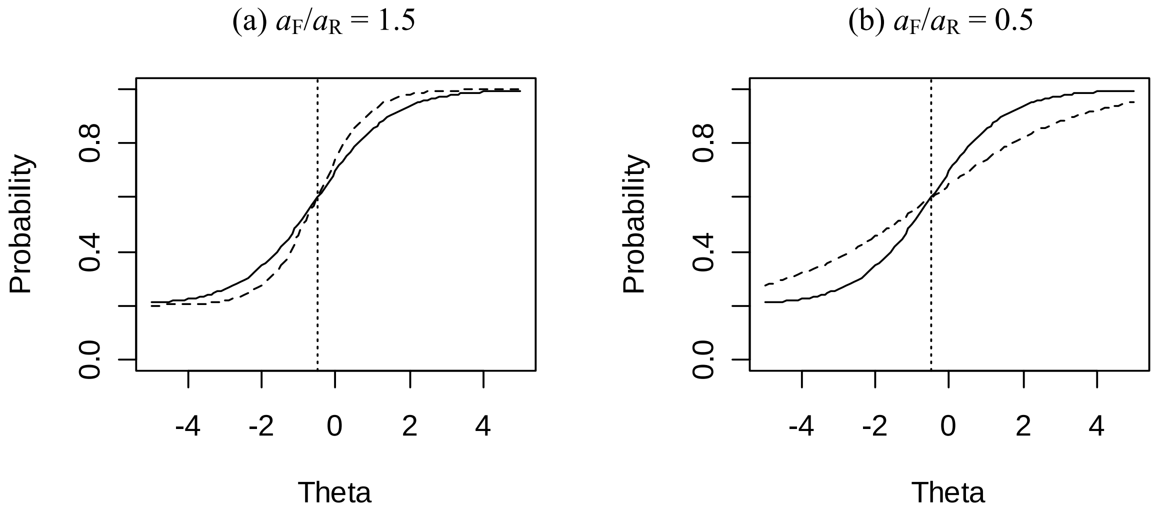
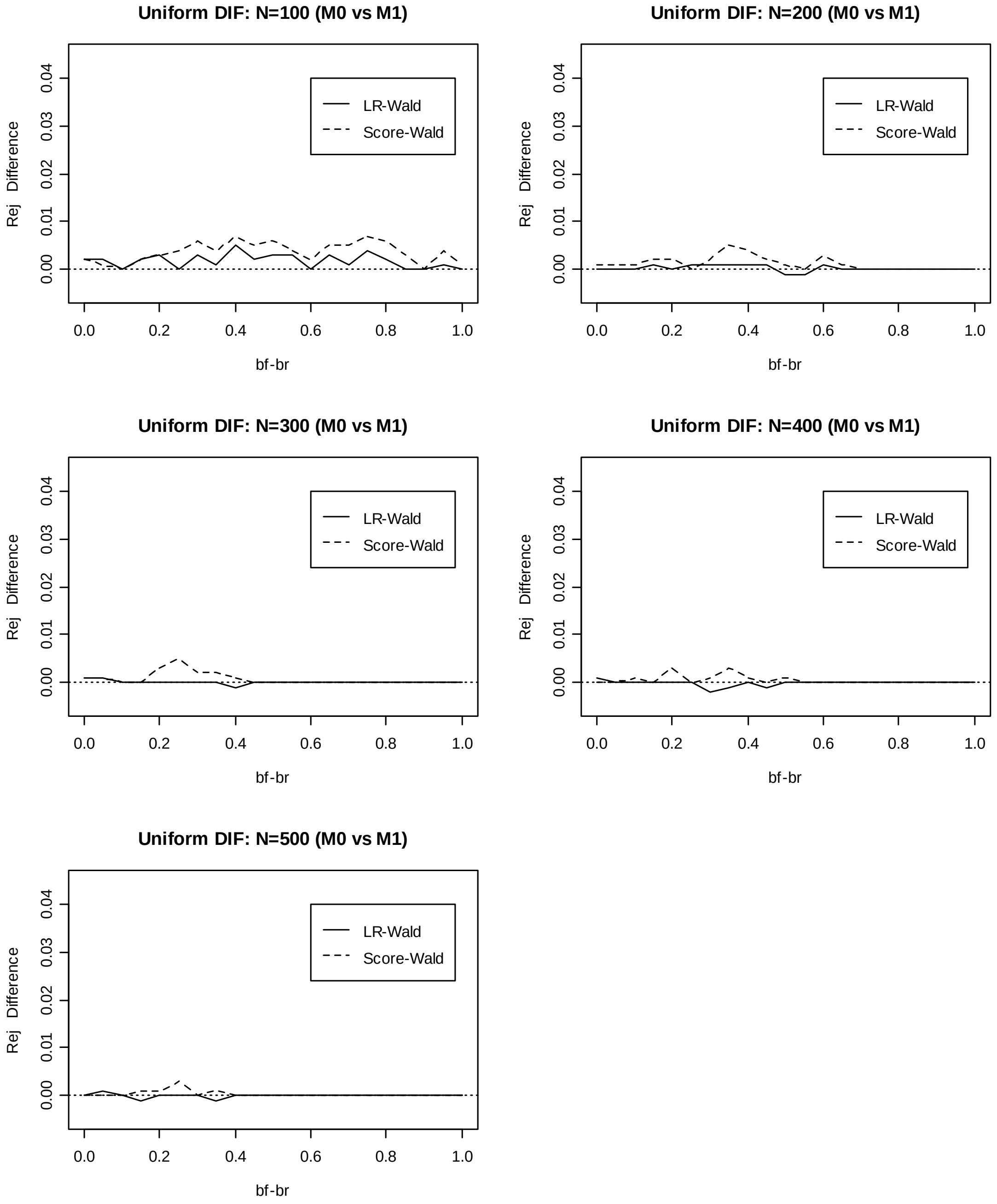


***Figure 3*. Item response function (IRF) differences for a_F/a_R = 1.5 and 0.5.**

Note. Dotted vertical lines represent item-difficulty points. Solid IRF lines have $a_R = 1$.



Uniform DIF: N=100 (M0 vs M1)

Uniform DIF: N=200 (M0 vs M1)

Uniform DIF: N=300 (M0 vs M1)

Uniform DIF: N=400 (M0 vs M1)

Uniform DIF: N=500 (M0 vs M1)

***Figure 4*. Differences in the three tests' rejection rates for detection of uniform differential item functioning (DIF) by targeted DIF testing.**

Note. Wald = the Wald test, LR = the likelihood-ratio test, Score = the score test, N = sample size per group, and bf – br = $b_F - b_R$. The plots show differences in the three tests' rejection rates when the Wald test is the baseline.

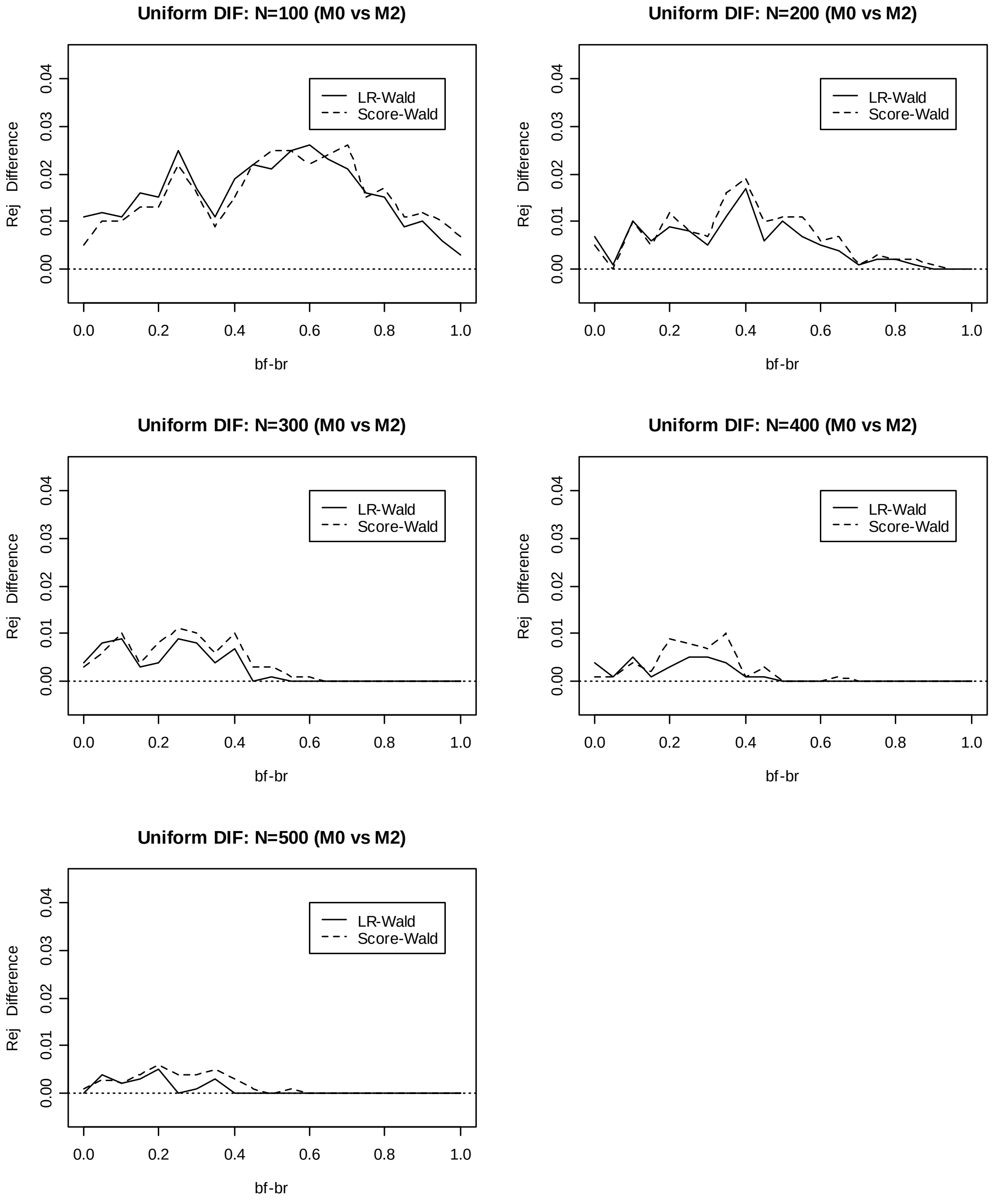


***Figure 5*. Differences in the three tests' rejection rates for detection of uniform differential item functioning (DIF) by general DIF testing.**

Note. Wald = the Wald test, LR = the likelihood-ratio test, Score = the score test, and bf – br = $b_F - b_R$. The plots show differences in the three tests' rejection rates when the Wald test is the baseline.

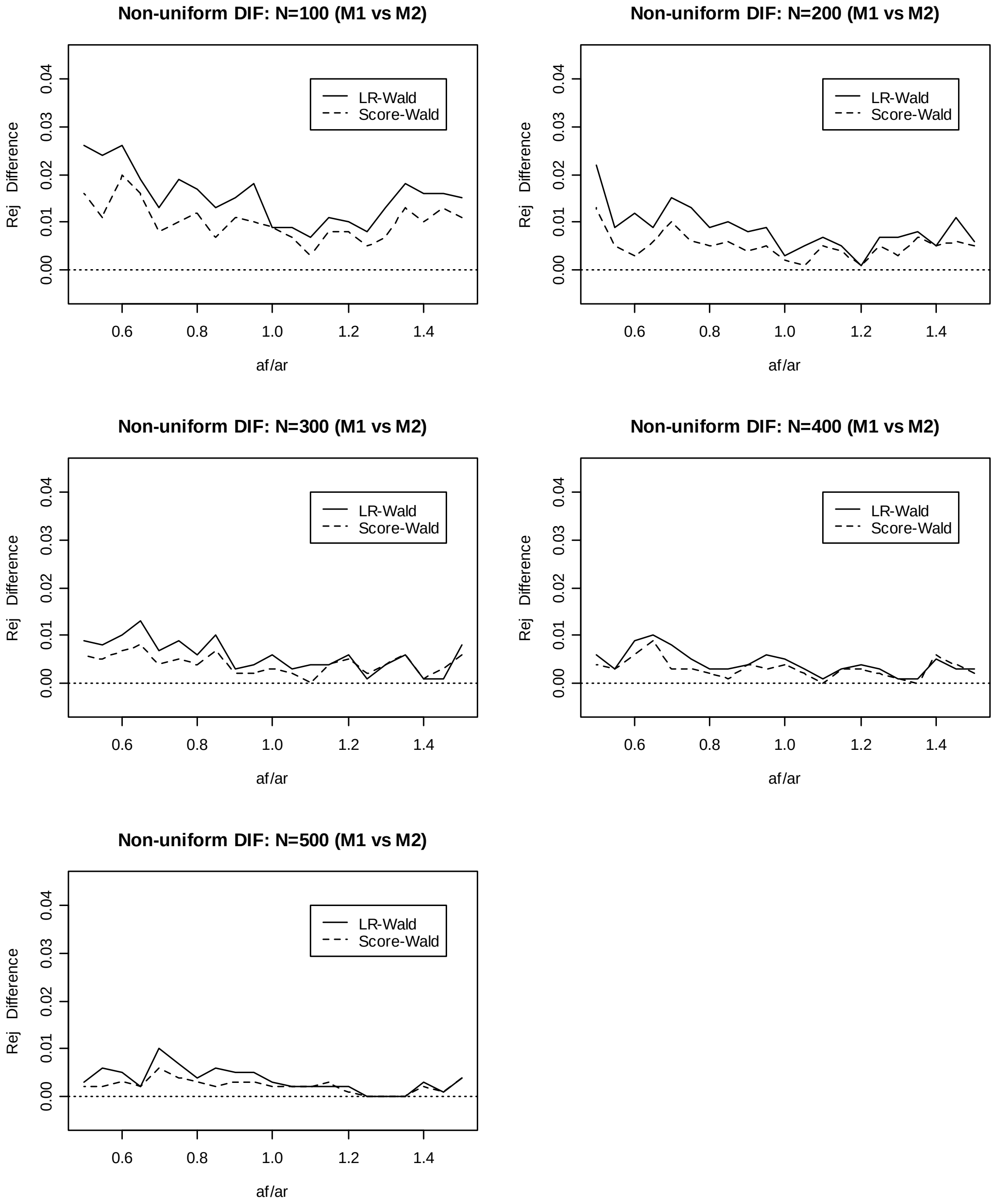


***Figure 6*. Differences in the three tests' rejection rates for detection of nonuniform differential item functioning (DIF) by targeted DIF testing.**

Note. Wald = the Wald test, LR = the likelihood-ratio test, Score = the score test, and af/ar = a_F/a_R. The plots show differences in the three tests' rejection rates when the Wald test is the baseline.

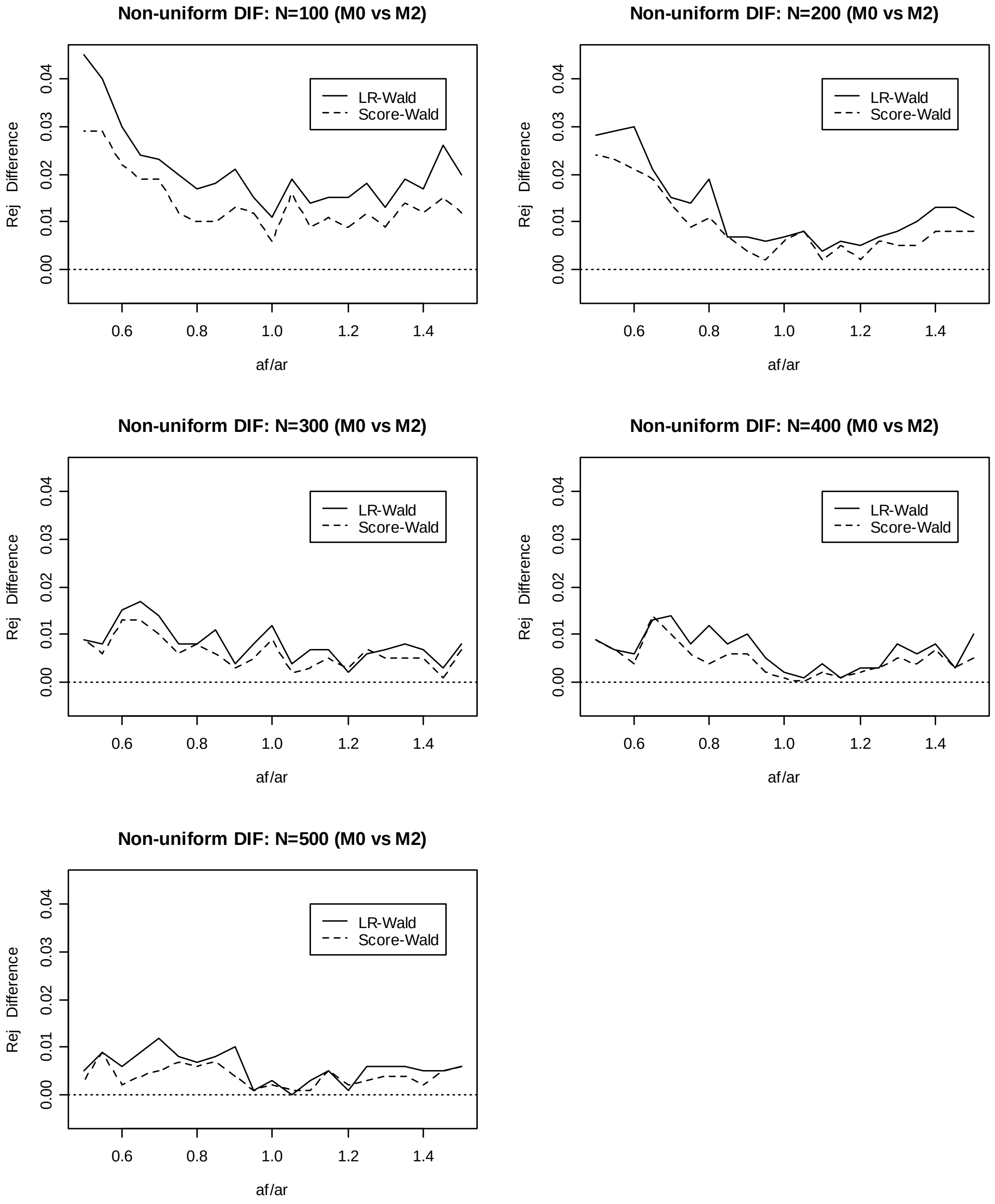


Figure 7. **Differences in the three tests' rejection rates for detection of nonuniform differential item functioning (DIF) by general DIF testing.**

Note. Wald = the Wald test, LR = the likelihood-ratio test, Score = the score test, and af/ar = a_F/a_R. The plots show differences in the three tests' rejection rates when the Wald test is the baseline.

Overall, the three tests behaved very similarly, and the differences in their rejection rates were small. The largest rejection-rate differences between the three tests occurred in about 3% of the replications for uniform DIF (Figures 4 and 5) and about 4.5% of those for nonuniform DIF (Figures 6 and 7). The differences decreased as both sample size and DIF increased. General DIF testing (M0 vs. M2) tended to create more differences than targeted testing. Nonuniform DIF showed more differences than uniform DIF. In general, the rejection rates of the score and LR tests were higher than those of the Wald test. For uniform DIF, the score test had the highest detection rate, the LR test the second highest, and the Wald test the lowest. For nonuniform DIF, the LR test showed the highest detection rate across different sample sizes and DIF sizes, the score test showed the second highest detection rate, and the Wald test showed the lowest. For both uniform and nonuniform DIF, the differences between the LR and score tests were smaller than the differences between the Wald test and either the LR or score test. Table 2 summarizes the average differences in rejection rates over all DIF magnitudes.

For targeted DIF testing, the average differences in rejection rate were very small. On average, they occurred in only 1.5% or less of the replications. For general DIF testing, differences occurred, on average, in up to 2.1% of the replications, but the same consistent patterns mentioned for Figures 4 through 7 can still be observed in Table 2.

Comparison of Targeted and General Differential-Item-Functioning Testing

Tables 3 and 4 show the differences in DIF detection rates between targeted and general DIF testing as percentages.

In Tables 3 and 4, positive numbers indicate a higher DIF detection rate for targeted DIF testing; negative numbers indicate a higher detection rate for general DIF testing. As the tables show, targeted DIF testing had a higher detection rate than general DIF testing, regardless of the test used (Wald, LR, or score). The row and column margins, which are averages, confirm that targeted DIF testing showed better performance than general DIF testing for detecting DIF. Across small and large sample sizes, targeted DIF testing showed detection rates 1–5% higher for uniform DIF and 1–4% higher for nonuniform DIF. Across different DIF magnitudes, targeted DIF testing showed detection rates 1–7% higher for uniform DIF and 1–8% higher for nonuniform DIF.

On average, for uniform DIF, the difference between targeted and general DIF testing decreased and became negligible as the sample size and $b_F - b_R$ increased (e.g., $N = 500$ with $b_F - b_R \geq 0.45$). For nonuniform DIF, larger sample sizes increased the difference between targeted and general DIF testing, especially when the focal-group slope was lower than the reference-group slope ($a_F/a_R < 1$).

Table 2

Average Differences in Rejection Rates Between the Wald, Likelihood Ratio (LR), and Score Tests

	Uniform DIF			
	M0 vs. M1		M0 vs. M2	
N	LR-Wald	Score-Wald	LR-Wald	Score-Wald
100	0.002	0.004	0.016	0.016
200	0.000	0.001	0.005	0.006
300	0.000	0.001	0.003	0.004
400	0.000	0.000	0.001	0.002
500	0.000	0.000	0.001	0.002
	Nonuniform DIF			
	M1 vs. M2		M0 vs. M2	
N	LR-Wald	Score-Wald	LR-Wald	Score-Wald
100	0.015	0.010	0.021	0.014
200	0.009	0.005	0.013	0.009
300	0.006	0.004	0.008	0.006
400	0.004	0.003	0.007	0.005
500	0.003	0.002	0.006	0.004

Note. The values were obtained by calculating averages over different DIF magnitudes.

Table 3

Percent Differences in Rejection Rates for Uniform Differential Item Functioning (DIF) between Targeted DIF and General DIF

	N = 100			*N* = 200			*N* = 300			*N* = 400			*N* = 500			
$b_F - b_R$	Wald	LR	Score	Wald	LR	Score	Wald	LR	Score	Wald	LR	Score	Wald	LR	Score	Average
0.00	2	1	1	0	-1	-1	0	0	0	-1	-2	-1	0	0	0	**0**
0.05	2	1	1	1	1	1	2	1	1	1	1	1	0	0	0	**1**
0.10	2	1	1	1	0	0	1	0	0	1	0	1	1	1	1	**1**
0.15	2	1	1	2	1	2	1	0	0	1	1	1	4	4	4	**2**
0.20	4	3	3	4	3	3	4	4	3	3	3	3	5	5	5	**4**
0.25	4	1	2	4	3	3	4	3	3	6	6	6	4	4	4	**4**
0.30	4	2	3	7	6	6	5	4	4	5	4	4	6	6	5	**5**
0.35	7	6	6	10	9	9	7	7	7	5	5	4	6	6	6	**7**
0.40	8	6	7	7	5	6	3	2	2	5	5	5	5	5	4	**5**
0.45	9	7	7	6	6	6	5	5	5	4	4	4	1	1	1	**5**
0.50	8	6	6	8	6	7	4	4	3	1	1	1	0	0	0	**4**
0.55	9	7	7	7	6	5	2	2	2	0	0	0	0	0	0	**3**
0.60	9	7	7	3	3	3	2	2	2	0	0	0	0	0	0	**3**
0.65	9	7	7	4	3	3	1	1	1	0	0	0	0	0	0	**2**
0.70	9	7	7	2	2	2	1	1	1	0	0	0	0	0	0	**2**
0.75	6	4	5	1	1	1	0	0	0	0	0	0	0	0	0	**1**
0.80	6	5	5	0	0	0	0	0	0	0	0	0	0	0	0	**1**
0.85	3	2	2	1	1	0	0	0	0	0	0	0	0	0	0	**1**
0.90	4	3	3	1	1	1	0	0	0	0	0	0	0	0	0	**1**
0.95	3	2	2	0	0	0	0	0	0	0	0	0	0	0	0	**1**
1.00	4	3	3	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Average	**5**	**4**	**4**	**3**	**3**	**3**	**2**	**2**	**2**	**2**	**1**	**1**	**2**	**2**	**1**	

Note. Wald = the Wald test, LR = the likelihood-ratio test, and Score = the score test. The values were calculated by 100 × (targeted DIF rejection rate – general DIF rejection rate) and rounded to an integer.

Table 4

Percent Differences in Rejection Rates for Nonuniform Differential Item Functioning (DIF) between Targeted DIF and General DIF

\	*N* = 100			*N* = 200			*N* = 300			*N* = 400			*N* = 500			
a_F/a_R	Wald	LR	Score	Wald	LR	Score	Wald	LR	Score	Wald	LR	Score	Wald	LR	Score	Average
1.00	0	0	0	-1	-1	-1	0	0	0	0	0	1	1	1	1	**0**
1.05	1	0	0	0	-1	-1	0	-1	0	-1	-1	-1	0	0	0	**-1**
0.95	0	0	0	2	2	2	1	1	1	1	1	1	1	2	2	**1**
1.10	0	-1	-1	1	1	1	1	0	0	-1	-1	-1	-1	-1	0	**0**
0.90	1	0	1	0	0	0	1	1	1	1	1	1	2	1	2	**1**
1.15	-1	-1	-1	0	0	0	0	0	0	1	1	1	1	1	1	**0**
0.85	0	0	0	3	4	3	2	2	2	4	4	4	3	2	2	**2**
1.20	-1	-1	-1	-1	-1	-1	0	0	0	2	2	2	-1	-1	-1	**0**
1.25	1	0	0	0	0	0	1	1	1	2	2	1	2	1	1	**1**
0.80	2	2	2	2	1	1	3	2	2	5	4	4	7	7	7	**3**
1.30	2	2	1	0	0	0	2	1	2	1	0	0	3	3	3	**1**
0.75	2	1	1	2	2	1	4	4	4	5	5	5	5	5	5	**3**
1.35	1	1	1	0	0	0	3	3	3	3	2	3	5	5	5	**2**
1.40	1	1	1	2	1	2	3	2	2	3	3	3	5	5	5	**3**
0.70	3	2	2	5	6	5	9	8	8	9	8	8	9	8	9	**7**
1.45	3	2	2	4	3	3	2	2	2	5	5	5	6	6	6	**4**
1.50	2	2	2	2	2	2	4	4	4	6	5	6	5	5	5	**4**
0.65	5	4	4	8	6	6	5	5	4	9	9	9	8	8	8	**7**
0.60	5	5	5	8	6	6	7	6	6	8	8	8	7	7	7	**7**
0.55	7	6	6	8	6	6	7	7	7	8	7	7	7	7	6	**7**
0.50	9	7	7	11	10	10	9	9	9	8	7	7	5	5	5	**8**
Average	**2**	**1**	**2**	**3**	**2**	**2**	**3**	**3**	**3**	**4**	**3**	**4**	**4**	**4**	**4**	

Note. Wald = the Wald test, LR = the likelihood-ratio test, and Score = the score test. The values were calculated by 100 × (targeted DIF rejection rate – general DIF rejection rate) and rounded to an integer.

Differential-Item-Functioning Detection Rate and Sample Size

Dorans and Holland (1993) used standardized *p* differences of 0.05 and 0.10 to describe items of negligible DIF, intermediate DIF (which warrants inspection), and large DIF (which is unusual and therefore warrants careful examination). For uniform DIF, I converted the item-difficulty difference, $b_F - b_R$, to the probability metric via the T1 statistic (see Wainer, 1993; p. 127, Equation 1), which is the average IRF difference weighted by the focal-group ability density function,

$$T1 = \int_{-\infty}^{\infty} \left(P_R\left(Y_i = 1|\theta\right) - P_F\left(Y_i = 1|\theta\right)\right) dG_F(\theta),$$

where $G_F(\theta)$ is the focal-group distribution function (Wainer used the notation of focal-group IRF minus reference-group IRF). T1 has the same form as the DIF estimator for the SIBTEST (Shealy & Stout, 1993), and it becomes the standardized *p*-difference index when the matching variable is an observed discrete variable. The $b_F - b_R$ values that most closely approximated 0.05 and 0.10 for uniform DIF were 0.20 (T1 = 0.045), 0.25 (T1 = 0.056), and 0.45 (T1 = 0.100). For nonuniform DIF, different item slope parameters ($a_F \neq a_R$)were converted to the probability metric, using an unsigned version of T1 (here called UT1),

$$UT1 = \int_{-\infty}^{\infty} \left| P_R\left(Y_i = 1|\theta\right) - P_F\left(Y_i = 1|\theta\right)\right| dG_F(\theta).$$

T1 and UT1 were evaluated using a numerical approximation with discrete θ points on [-5,5] in increments of .05. The a_F/a_R values closest to 0.05 and 0.10 were 1.50 (UT1 = 0.048), 0.65 (UT1 = 0.052), and 0.50 (UT1 = 0.080). UT1 differs from T1 in that it uses the absolute IRF difference rather than just the IRF difference. However, as an unsigned measure, UT1 can still be interpreted as the expected score difference at the item level as an unsigned measure. Table 5 shows the selected simulation conditions and the rejection rates whose T1 and UT1 values were closest to 0.05 and 0.10.

Table 5

Differential Item Functioning (DIF) Detection Rates and Sample Size for Intermediate and Large DIF

Uniform DIF																
Targeted DIF (M0 vs. M1)		*N* = 100			*N* = 200			*N* = 300			*N* = 400			*N* = 500		
$b_F - b_R$	T1	Wald	LR	Score	Wald	LR	Score	Wald	LR	Score	Wald	LR	Score	Wald	LR	Score
0.20	0.045	0.124	0.127	0.127	0.191	0.191	0.193	0.261	0.261	0.264	0.350	0.350	0.353	0.435	0.435	0.436
0.25	0.056	0.170	0.170	0.174	0.296	0.297	0.296	0.387	0.387	0.392	0.464	0.464	0.464	0.550	0.550	0.553
0.45	0.100	0.386	0.388	0.391	0.649	0.650	0.651	0.796	0.796	0.796	0.903	0.902	0.903	0.951	0.951	0.951
Uniform DIF																
General DIF (M0 vs. M2)																
$b_F - b_R$	T1															
0.20	0.045	0.086	0.101	0.099	0.154	0.163	0.166	0.222	0.226	0.230	0.317	0.320	0.326	0.383	0.388	0.389
0.25	0.056	0.132	0.157	0.154	0.254	0.262	0.262	0.352	0.361	0.363	0.399	0.404	0.407	0.507	0.507	0.511
0.45	0.100	0.298	0.320	0.320	0.586	0.592	0.596	0.748	0.748	0.751	0.860	0.861	0.863	0.943	0.943	0.944
Nonuniform DIF																
Targeted DIF (M1 vs. M2)																
a_F/a_R	UT1															
1.50	0.048	0.078	0.093	0.089	0.135	0.141	0.140	0.167	0.175	0.173	0.213	0.216	0.215	0.249	0.253	0.253
0.65	0.052	0.140	0.159	0.156	0.267	0.276	0.273	0.348	0.361	0.356	0.486	0.496	0.495	0.582	0.584	0.584
0.50	0.080	0.289	0.315	0.305	0.521	0.543	0.534	0.701	0.710	0.707	0.830	0.836	0.834	0.891	0.894	0.893
Nonuniform DIF																
General DIF (M0 vs. M2)																
a_F/a_R	UT1															
1.50	0.048	0.058	0.078	0.070	0.112	0.123	0.120	0.130	0.138	0.137	0.151	0.161	0.156	0.194	0.200	0.200
0.65	0.052	0.092	0.116	0.111	0.192	0.213	0.211	0.298	0.315	0.311	0.392	0.405	0.406	0.498	0.507	0.502
0.50	0.080	0.201	0.246	0.230	0.415	0.443	0.439	0.611	0.620	0.620	0.753	0.762	0.762	0.838	0.843	0.841

Note. Wald = the Wald test, LR = the likelihood-ratio test, and Score = the score test. T1 and UT1 values were rounded up to the third decimal place.

For intermediate DIF items (T1 ≈ 0.5; UT1 ≈ 0.5), the detection rates of targeted DIF testing ranged from about 8% to about 58% of the replications as the sample size increased from 100 to 500 per group; the detection rates of general DIF testing tended to be lower, ranging from about 6% to about 51% of the replications. When an item had relatively large DIF (T1 = 0.1 or UT1 = 0.08), the detection rates across different sample sizes were 29–95% of the replications for targeted DIF testing and 20–94% of the replications for general DIF testing.

With a sample size of 200 per group, the best detection rate is about 30% for an intermediate uniform DIF item, about 65% for a large uniform DIF item, and about 54% for a nonuniform DIF item with a UT1 of 0.8. When a DIF detection rate of more than 70% is desired for a large DIF item, a sample size of at least 300 per group appears to be a good choice with any test (Wald, LR, or score) and any DIF testing mode (targeted or general). Under this study's item parameters, test length and group-ability conditions, if a sample size of 500 per group is used, a medium DIF item will be detected at least 51% of the time, and a large uniform DIF item will be detected at least 94% of the time, whatever test and DIF testing mode are used.

Summary and Discussion

At least three statistical hypothesis tests can be used in logistic regression applications: the Wald, LR, and score tests. All three have asymptotic chi-square sampling distributions, but they can yield different results with finite samples. Swaminathan and Rogers (1990) introduced use of the Wald test. Although well-known hypothesis tests, the LR test and especially the score tests have not been popular, and their performance in logistic-regression DIF detection has not been rigorously compared. In addition, it does not seem to be well-recognized that the score-test statistic is equivalent to the MH chi-square test (Day & Byar, 1979). The study results showed that overall, the three tests behaved very similarly; differences in DIF detection rates were small. Especially with targeted uniform DIF testing, the three tests performed virtually the same. When the sample size and DIF magnitude became large, the differences between the tests became negligible.

A few consistent patterns, however, should be addressed. The tests' DIF-detection rates differed more with general DIF testing than with targeted DIF testing and were lower with general DIF testing, regardless of which test was used. Also, DIF-detection rates differed more with nonuniform DIF than with uniform DIF. The LR and score tests had better DIF-detection rates than the Wald test across different test modes (targeted vs. general), DIF types, and sample sizes. The score test was the best for detecting uniform DIF, and the LR test was the best for detecting nonuniform DIF. The author recommends either the LR or score test whenever either is available, especially when the sample size is not large. The slightly but consistently higher DIF detection rates of the LR and score tests appear to be in accordance with previous findings on the Wald test's statistical power (e.g., Fears, Benichou, & Gail, 1996; Hauck & Donner, 1977; Pawitan, 2000).

Targeted DIF testing performed better than general DIF testing. To make full use of logistic regression for any type of DIF detection, practitioners can take the following steps (presented as a flowchart in Figure 8):

1. Conduct targeted DIF testing for nonuniform DIF (M1 vs. M2).
2. If the result of Step 1 is not statistically significant, conduct targeted DIF testing for uniform DIF (M0 vs. M1). If the result of Step 1 *is* statistically significant, stop the statistical DIF testing and conclude statistically significant nonuniform DIF.
3. If you conducted targeted DIF testing for uniform DIF and the result was statistically significant, conclude statistically significant uniform DIF.. If you conducted targeted DIF testing for uniform DIF and the result was *not* statistically significant, stop the statistical DIF testing and conclude noDIF..

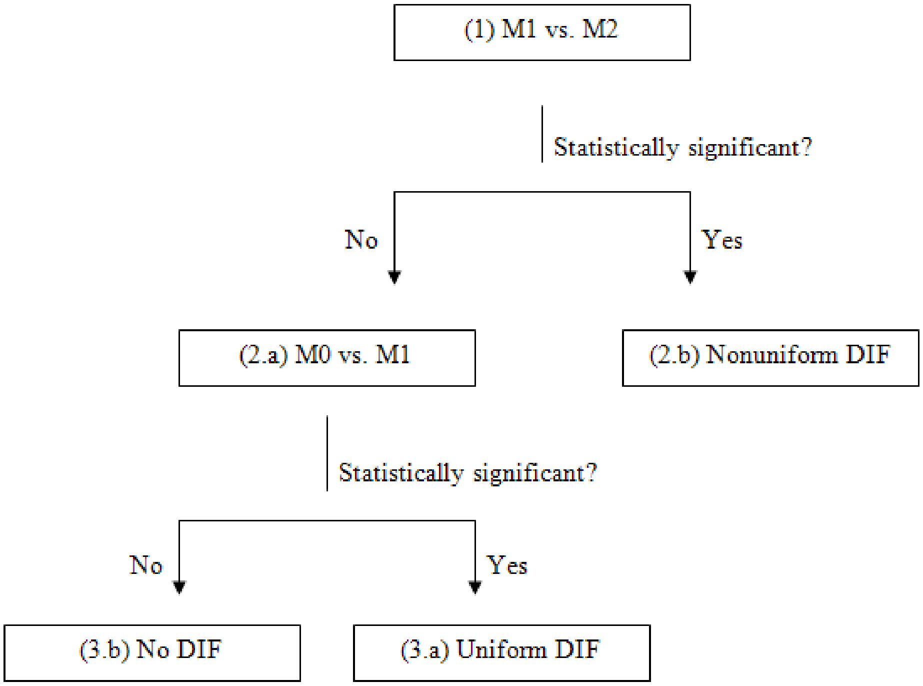


Figure 8**. Strategy for detection of differential item functioning (DIF) with targeted DIF testing.**

In practice, detecting large DIF is *ge*nerally of the utmost importance. The current study findings seem to suggest that a sample size of 100 per group is too small to detect large DIF frequently (see Table 5 for the detection rate when T1 = 1.0). Given the simulation conditions used in this study, a sample size of at least 200 or higher appear to be recommended to detect large (uniform or nonuniform) DIF at least half or more than half of the time. With a minimum of 300 per group, this study shows that an item with large DIF would be detected at least 60% of the time, whatever testing mode or test was used. For a sample of 300 per group, the uniform DIF detection rate for an item with large DIF went up to 75% in general DIF testing and 80% in targeted DIF testing (see Table 5). Note also that differences in DIF detection rates between the three tests become more inconsequential with a sample of 300 per group than with samples of 100 or 200 per group (see Figures 5 and 7).

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement*, *43*, 313-333.

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

Cochran, W. G. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics, 10*, 417–451.

Cox. D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. New York: Chapman & Hall/CRC.

Day, N. E., & Byar, D. P. (1979). Testing hypotheses in case-control studies: Equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics, 35*, 623–630.

Dobson, A. J. (2002). *An introduction to generalized linear models* (2nd ed.). New York: Chapman & Hall/CRC.

Dorans, N., & 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., & Potenza, M. (1994). *Equity assessment for polytomously scored items: A taxonomy of procedures for assessing differential item functioning* (ETS Research Rep. No. RR-94-99). Princeton, NJ: ETS.

Fears, T. R., Benichou, J., & Gail, M. H. (1996). A reminder of the fallibility of the Wald statistic. *The American Statistician*, *50,* 226–227.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8*, 647–667.

Harrell, F. E., Jr. (2001). *Regression modeling strategies: With applications to linear models, logistics regression, and survival analysis*. New York: Springer.

Hauck, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association, 72*, 851–853.

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

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

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: Swets & Zitlinger.

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 National Cancer Institution, 22*, 719–768.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York: Chapman & Hall.

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

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*, 92–109.

Pawitan, Y. (2000). A reminder of the fallibility of the Wald statistic: Likelihood explanation. *The American Statistician*, *54,* 54–56.

Pregibon, D. (1982). Score tests in GLIM with applications. In R. Gilchrist (Ed.), *Lecture notes in statistics: No. 14. GLIM 82: Proceedings of the international conference on generalized linear models* (pp. 87–97). New York: Springer.

The R project for statistical computing. (2002). Retrieved November 5, 2008, from http://www.r-project.org.

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

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197–207.

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

Spray, J. A., & Miller, T. R. (1992). Performance of the Mantel-Haenszel statistic and the standardized difference in proportions correct when population ability distributions are incongruent (ACT Research Rep. Series No. 92-1). Iowa City, IA: ACT.

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

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

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

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

Vaeth, M. (1985). On the use of Wald's test in exponential families. *International Statistical Review, 53*, 199–214.

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

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

Notes

[1] In this research, uniform DIF is defined as parallel three-parameter logistic (3PL) model item response function (IRF) differences (the same slope and guessing parameters but different difficulty parameters); nonuniform DIF is defined as nonparallel (crossing) 3PL model IRF differences (the same difficulty and guessing parameters but different slope parameters).

[2] The result of targeted DIF testing (M0 vs. M1 or M1 vs. M2) does not of itself indicate that there is no statistically significant DIF. If the comparison of M1 and M2 is not statistically significant, there may nevertheless be statistically significant uniform DIF (by comparing M0 and M1). Also, even if the comparison of M0 and M1 does not indicate statistically significant uniform DIF, there may be statistically significant nonuniform DIF (by comparing M1 and M2). Therefore, M0-versus-M1 comparisons and M1-versus-M2 comparisons are referred to as *targeted*. They require an a priori hypothesis—assumption of uniform DIF or nonuniform DIF—to allow a conclusion of no DIF.

[3] Although nonuniform DIF due to different discriminations may not occur as often as uniform DIF, this study used the different discrimination as an approximation of one possible crossing IRF DIF scenario. Bolt and Gierl (2006) mentioned that translation DIF studies have shown a wide variety of DIF forms, including nonuniform (crossing) DIF. Also, their real data analysis for DIF showed some examples of crossing DIF items caused by discrimination differences.

[4] Data were generated by the 3PL model, which has a nonzero lower asymptote (the mean guessing-parameter value was 0.19, and the studied-item guessing-parameter value was 0.2). Note, however, that the lower asymptote for the logistic regression is zero.

[5] Although the differences from no DIF ($a_F = a_R = 1$) are the same for the two largest discrimination differences ($a_F/a_R = 0.5$ and 1.5), the IRFs were crossed such that $a_F/a_R = 0.5$ formed a larger IRF area difference than $a_F/a_R = 1.5$. The unsigned areas between IRFs calculated by Raju’s (1988, 1990) formula gave 0.65 for $a_F/a_R = 0.5$ and 0.22 for $a_F/a_R =1.5$.