DOCUMENT RESUME

ED 469 180 TM 034 485

AUTHOR	Schnipke, Deborah L.; Roussos, Louis A.; Pashley, Peter J.
TITLE	A Comparison of Mantel-Haenszel Differential Item Functioning Parameters. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-RR-98-03
PUB DATE	2000-09-00
NOTE	17p.; For a related report on the Mantel-Haenszel Differential Item Functioning Parameters, see TM 034 484.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	College Entrance Examinations; Comparative Analysis; *Item Bias; Item Response Theory; Law Schools; Simulation; *Test Items
IDENTIFIERS	Item Parameters; *Law School Admission Test; *Mantel Haenszel Procedure

ABSTRACT

Differential item functioning (DIF) analyses are conducted to investigate how items function in various subgroups. The Mantel-Haenszel (MH) DIF statistic is used at the Law School Admission Council and other testing companies. When item functioning can be well-described in terms of a one- or two-parameter logistic item response theory (IRT) model and subgroup differences can be conceptualized as a difference in item difficulty, the MH DIF statistic can be readily understood in terms of the IRT item parameters. When items follow the three-parameter logistic IRT model, however, the relationship between MH DIF statistic and IRT is more complicated, and several competing parameters that relate the statistic to IRT parameters have been proposed. The goal of this study was to investigate various MH DIF parameters to determine which is most appropriate. The parameters were compared with the MH DIF statistic using both simulated and real data from a recent administration of the Law School Admission Test (responses of 20,092 white male test takers). Results suggest that the most appropriate parameter is the one that is theoretically most similar to the MH DIF statistic itself. (Contains 3 figures, 2 tables, and 16 references.) (Author/SLD)

LSAC RESEARCH REPORT SERIES

TM

A Comparison of Mantel-Haenszel Differential Item Functioning Parameters

Deborah L. Schnipke, Louis A. Roussos,
and Peter J. Pashley
Law School Admission Council

Law School Admission Council
Research Report 98-03
September 2000

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 197 law schools in the United States and Canada.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

Items on large-scale standardized tests, such as the Law School Admission Test (LSAT), undergo an extensive sensitivity review before they are ever presented to test takers. Despite precautions, some items may still function differently among subgroups, so statistical analyses of differential item functioning (DIF) are performed after test takers respond to the items. Many DIF procedures have been developed, but the Mantel-Haenszel (MH) is the primary DIF procedure used at the Law School Admission Council (LSAC) and other major testing companies.

The MH procedure was first proposed for situations in which items cannot be answered correctly by guessing. Under this constraint, the MH statistic has a direct relationship to item difficulty, as specified by item response theory, so the statistic's behavior and interpretation are well understood. When items can be answered correctly by guessing (e.g., many multiple-choice items), the relationship between the MH DIF statistic and IRT difficulty is more complicated, so the behavior and interpretation of the statistic are not well understood. Several theorists have proposed MH DIF parameters in the attempt to explain the statistic's behavior under these more complicated circumstances. The purpose of the present study is to compare the proposed MH DIF parameters in order to determine which parameter most accurately captures the MH DIF statistic's behavior.

Three MH DIF parameters were compared with values of the MH DIF statistic in simulated and real data. Not surprisingly, of the three parameters investigated, the one that is most theoretically similar to the MH DIF statistic itself was found to best explain the statistic's behavior under a variety of conditions.

Abstract

Differential item functioning (DIF) analyses are conducted to investigate how items function in various subgroups. The Mantel-Haenszel (MH) DIF statistic is used at the Law School Admission Council and other testing companies. When item functioning can be well-described in terms of a one- or two-parameter logistic item response theory (IRT) model and subgroup differences can be conceptualized as a difference in item difficulty, the MH DIF statistic can be readily understood in terms of the IRT item parameters. When items follow the three-parameter logistic IRT model, however, the relationship between the MH DIF statistic and IRT is more complicated, and several competing parameters that relate the statistic to IRT parameters have been proposed. The goal of the present paper is to investigate various MH DIF parameters to determine which is most appropriate. The parameters are compared with the MH DIF statistic using both simulated and real data. Results suggest that the most appropriate parameter is the one that is theoretically most similar to the MH DIF statistic itself.

Introduction

Items on large-scale standardized tests, such as the Law School Admission Test (LSAT), undergo an extensive sensitivity review before they are ever presented to test takers. Despite precautions, some items may still function differently among subgroups, so statistical analyses of differential item functioning (DIF) are performed after test takers respond to the items. Many DIF procedures have been developed, but the Mantel-Haenszel procedure (MH; Mantel & Haenszel, 1959), as modified by Holland and Thayer (1988), is the primary DIF procedure used at the Law School Admission Council (LSAC) and other major testing companies.

The MH procedure calculates a scaled natural logarithm of a weighted average of the ratio of the odds of a correct response for two subgroups of interest. The MH procedure was first proposed for situations in which items cannot be answered correctly by guessing. Under this constraint, the MH statistic has a direct relationship to item difficulty, as specified by item response theory (IRT), so the statistic's behavior and interpretation are

well understood. Accordingly, the MH statistic is often interpreted as a scaled difference in item difficulty for the two subgroups (e.g., Donoghue, Holland, & Thayer, 1993; Zwick, Thayer, & Wingersky, 1994, 1995).

When items can be answered correctly by guessing (e.g., many multiple-choice items), the relationship between the MH DIF statistic and IRT difficulty is more complicated, so the behavior and interpretation of the statistic are not well understood. Several MH DIF parameters have been proposed in the attempt to explain the statistic's behavior under these more complicated circumstances (Donoghue, Holland, & Thayer, 1993; Roussos, Schnipke, & Pashley, in press; Spray & Miller, 1992; Zwick, Thayer, & Lewis, 1997; Zwick, Thayer, & Wingersky, 1994). The purpose of the present study is to compare the proposed MH DIF parameters in order to determine which parameter most accurately captures the MH DIF statistic's behavior. Comparisons are made using both simulated and real data, and the MH DIF parameters were computed on the basis of IRT item parameters (true parameters for the simulated data, and parameter estimates for the real data).

Much of the data for this paper was taken from Roussos, Schnipke, and Pashley (in press). We would like to thank David J. Scrams for his extensive comments on an earlier draft of this paper, Lisa Anthony for compiling the data for the real-data analysis, and Lynda Reese for reviewing the paper.

MH DIF Statistic

The MH DIF statistic, $\hat{\Delta}$, compares the odds of a correct response for two subgroups after controlling for differences in overall ability. The groups are often called the *reference* and *focal* groups, and if an item functions similarly for the groups, their odds of a correct response should be equal after controlling for differences in overall ability. Differences in overall ability are controlled by comparing subgroups separately for each total score, then aggregating the comparisons across scores. Specifically, the empirical odds (the number of test takers who answered the item correctly, divided by the number of test takers who answered incorrectly) for each subgroup is calculated separately for test takers receiving each total score. A ratio of the empirical odds for the two groups is calculated for each total score, and these odds ratios (weighted according to their statistical stability) are then averaged. The final statistic is a scaled natural logarithm of the resulting averaged ratio so that the value has certain desirable statistical properties.

Mathematically, the empirical odds of a correct response for the reference group is given by C_{Rs}/I_{Rs}, where C_{Rs} is the number of test takers in the reference group who received a total score of s and answered the item correctly, and I_{Rs} is the number who received a total score of s but answered incorrectly. Similarly, the empirical odds of a correct response for the focal group is given by C_{Fs}/I_{Fs}. The empirical score-level odds ratio, $\hat{\alpha}_s$, is given by

$$\hat{\alpha}_s = \frac{C_{Rs}/I_{Rs}}{C_{Fs}/I_{Fs}} = \frac{C_{Rs}I_{Fs}}{C_{Fs}I_{Rs}}. \qquad (1)$$

The score-level odds ratio ranges from 0 to ∞. If the two groups have the same odds of a correct response (no DIF) at score level s, the odds ratio will be 1. If the two groups do not have the same odds of a correct response (DIF) at score level s, the odds ratio will not be 1. The score-level odds ratios are calculated at each score level, and they are combined for an overall measure of the difference in odds of a correct response for the two groups, regardless of score.

The score-level odds ratios are assumed to be constant across score level, so each score-level odds ratio is assumed to be estimating the same overall odds ratio. Therefore, any weighted average of the score-level odds ratios may be used. If C_{Fs} or I_{Rs} is close to 0, $\hat{\alpha}_s$ will be unstable because these terms are in the denominator of $\hat{\alpha}_s$. Thus it is reasonable to choose weights that minimize unstable values of $\hat{\alpha}_s$. Mantel and Haenszel (1959) proposed weights that do just that. The MH odds ratio, $\hat{\alpha}$, is given by

$$\hat{\alpha} = \frac{\sum\left(\frac{C_{Fs} I_{Rs}}{N_{Total,s}} \hat{\alpha}_s\right)}{\sum\left(\frac{C_{Fs} I_{Rs}}{N_{Total,s}}\right)}, \quad (2)$$

where $\hat{\alpha}_s$ is given by Equation 1, $N_{Total,s}$ is the number of test takers who received a score of s, and the summations are across score levels. When the odds of a correct response are the same for the two groups, regardless of score, $\hat{\alpha} = 1$. Otherwise, $\hat{\alpha} \neq 1$ indicates that the odds are not the same (i.e., the item contains DIF).

The MH odds ratio, $\hat{\alpha}$, which ranges from 0 to ∞, is not on an intuitive scale. Holland and Thayer (1988) transformed the MH odds ratio to the Educational Testing Service (ETS) "delta scale," defining the MH DIF statistic, $\hat{\Delta}$, as

$$\hat{\Delta} = -2.35 \ln(\hat{\alpha}). \quad (3)$$

The delta scale is an inverse normal transformation of the percent correct on the item to a linear scale with a mean of 13 and a standard deviation of 4. Test developers at ETS use this scale as a measure of item difficulty. Whereas $\hat{\alpha}$ is interpreted in terms of the odds of answering an item correctly, the MH DIF statistic, $\hat{\Delta}$, is interpreted as a difference in item difficulty for reference- and focal-group test takers on the ETS delta scale (Holland & Thayer, 1988).

ETS developed a classification scheme to help determine when to flag items for moderate and large DIF (e.g., Zieky, 1993). If $|\hat{\Delta}| < 1$ or $\hat{\Delta}$ is not significantly larger than 0, then the item is considered to contain no detectable DIF and is given an "A" flag. If $|\hat{\Delta}| > 1.5$ and $\hat{\Delta}$ is significantly greater than 1, the item is considered to contain large DIF and is given a "C" flag. All other items are given a "B" flag and are considered to contain moderate DIF. These classifications are commonly used at LSAC and other testing companies to identify items for further review.

In the next section, several proposed MH DIF parameters will be reviewed. The parameters can all be used with IRT to define the amount of "true" DIF in an item. Several of the parameters are quite general and can use any model that relates the probability of a correct response to item and test-taker characteristics.

MH DIF Parameters

Holland and Thayer (1988) showed that $\hat{\Delta}$ has a direct theoretical relationship to item difficulty, as defined by item response theory (IRT), when responses follow the one-parameter logistic (1PL) IRT model. In particular, they showed that in the 1PL case, the population odds ratio, α, is given by $\alpha_{1PL} = e^{b_R - b_F}$, where b_R and b_F are the IRT difficulty values for the reference and focal groups, respectively. Thus, in the 1PL case, the MH Δ value is proportional to the difference in b values for the two subgroups.

Donoghue, Holland, and Thayer (1993) expanded Holland and Thayer's (1988) work and determined a more general relationship between DIF, as measured by the MH procedure, and the IRT definition of DIF (as a difference in item difficulty [b values] for the two groups) when there is no guessing ($c = 0$ for all items). Specifically, they determined that when (1) the Rasch model holds, (2) the matching variable is the number-right

score based on all items including the studied item,[1] and (3) none of the items in the matching variable have IRT DIF, except possibly the item being tested for DIF, then

$$\Delta_{DHT} = 4a(b_R - b_F), \tag{4}$$

where a is constant for all items in the analysis.

Zwick, Thayer, and Wingersky (1994) noted that $4a(b_R - b_F)$ generally overestimates the amount of DIF detected by $\hat{\Delta}$ when data follow the 3PL IRT model (which breaks the assumption that the Rasch model holds). They determined empirically that

$$\Delta_{ZTW} \approx 3a(b_R - b_F) \tag{5}$$

more accurately reflects $\hat{\Delta}$ in the 3PL case.

Zwick, Thayer, and Lewis (1997) defined the MH DIF parameter in a way that is more similar to the MH DIF statistic. They defined the score-level odds ratios in terms of theoretical probabilities of answering the item correctly or incorrectly, rather than the empirical numbers of test takers answering correctly or incorrectly. They weighted the score-level odds ratios by the ability distribution of the reference group [$f_R(\theta)$], integrated across ability, then took the natural logarithm and multiplied by –2.35, as is done with the MH DIF statistic. Their formula is given by

$$\Delta_{ZTL} = -2.35 \ln \int f_R(\theta)\alpha(\theta)d\theta, \tag{6}$$

where $\alpha(\theta)$ is the theoretical score-level odds ratio and is given by

$$\alpha(\theta) = \frac{P_R(\theta)Q_F(\theta)}{P_F(\theta)Q_R(\theta)}, \tag{7}$$

where $P_R(\theta)$ is the probability of a correct response for the reference group, $Q_R(\theta)$ is the probability of an incorrect response for the reference group, and $P_F(\theta)$ and $Q_F(\theta)$ are similarly defined for the focal group.

Spray and Miller (1992) defined the MH DIF parameter in the 3PL case by taking the MH DIF statistic and substituting θ (IRT ability value) for s (number-right score). They then determined the theoretical form of the MH DIF parameter that was most logical and consistent with the definition of the MH DIF statistic. Roussos, Schnipke, and Pashley (in press) also started with the MH DIF statistic to determine what parameter the MH DIF statistic estimates. They applied asymptotic theory to the MH DIF statistic (letting the number of test takers and items become infinitely large) and derived the same formula that Spray and Miller (1992) proposed. The Spray and Miller/Roussos, Schnipke, and Pashley MH DIF parameter is given by

$$\Delta_{SM/RSP} = -2.35 \ln \left[\frac{\int_{-\infty}^{\infty} \left[P_F(\theta)Q_R(\theta) \frac{f_R(\theta)f_F(\theta)}{\gamma_R f_R(\theta) + \gamma_F f_F(\theta)} \alpha(\theta) \right] d\theta}{\int_{-\infty}^{\infty} \left[P_F(\theta)Q_R(\theta) \frac{f_R(\theta)f_F(\theta)}{\gamma_R f_R(\theta) + \gamma_F f_F(\theta)} \right] d\theta} \right], \tag{8}$$

[1] The "studied item" is the item being tested for DIF.

where $\alpha(\theta)$ is defined in Equation 7, γ_R and γ_F are the proportions of test takers in the reference and focal groups, respectively, $f_F(\theta)$ is the ability distribution of the focal group, and the other terms are defined as before.

The only difference between Equations 6 and 8 is the weighting function. Zwick, Thayer, and Lewis (1997) used the reference group ability distribution as the weights (Equation 6), whereas Spray and Miller (1992) and Roussos, Schnipke, and Pashley (in press) used weights that are theoretically comparable to the weights used by the MH DIF statistic, $\hat{\Delta}$.

Note that if the 2PL IRT formula is substituted into the ability-level odds ratio (Equation 7), the result is

$$\alpha_{2PL}(\theta) = e^{-1.7a(b_R - b_F)}, \tag{9}$$

which does not depend on θ, as assumed by the MH procedure. Substituting this term into either the Zwick, Thayer, and Lewis (1997) formula (Equation 6) or the Spray and Miller (1992), Roussos, Schnipke, and Pashley (in press) formula (Equation 8) results in the weights dropping out because $\alpha_{2PL}(\theta)$ comes outside of the integral. Thus, in the 2PL case, both formulas reduce to

$$\Delta_{2PL} = -2.35 \ln\left(e^{-1.7a(b_R - b_F)}\right) = 4a(b_R - b_F), \tag{10}$$

identical to Equation 4, but without the equal-discrimination constraint.

If the 3PL IRT model is substituted into the ability-level odds ratio (Equation 7), the result is

$$\alpha_{3PL}(\theta) = \left[\frac{1 + ce^{-1.7a(\theta - b_R)}}{1 + ce^{-1.7a(\theta - b_F)}}\right] e^{-1.7a(b_R - b_F)}, \tag{11}$$

which does depend on θ, breaking an assumption of the MH procedure. The Zwick, Thayer, and Lewis (1997) formula (Equation 6) and the Spray and Miller (1992), Roussos, Schnipke, and Pashley (in press) formula (Equation 8) will result in different values because the weights do not drop out. Furthermore, these values will not be the same as the corresponding 2PL values (Equation 10) in the 3PL case (i.e., when $c \neq 0$). Thus the important comparisons for the various MH DIF parameters are in the 3PL case. The next section will evaluate the various MH DIF parameters in terms of the behavior of the MH DIF statistic, $\hat{\Delta}$, using simulated 3PL data. The section after that will evaluate the various MH DIF parameters in terms of the behavior of the MH DIF statistic, $\hat{\Delta}$, using empirical data.

Comparison of Δs With $\hat{\Delta}$ in Simulated Data

The three MH DIF parameters ($\Delta_{SM/RSP}$, Δ_{ZTL}, and Δ_{2PL}) were compared to $\hat{\Delta}$ values which were reported in Allen and Donoghue (1996). The conditions used by Allen and Donoghue are described next.

Method

Allen and Donoghue (1996) reported $\hat{\Delta}$ values for 15 simulated items. In the simulation study, three values of a were used: .5, 1, and 1.5. Five values of b_R were used: –2, –1, 0, 1, and 2. For all items, c = .2 and b_F = b_R +.4. Additionally, $\theta_R \sim N(0, .7)$ and $\theta_F \sim N(-.7, .8)$. The reference-group size was 5,100, and the focal-group

size was 1,050. To provide an example of the magnitude of the IRT DIF that was induced, the item characteristic curves (ICCs) for one item with DIF are shown in Figure 1. A separate ICC is provided for each group.

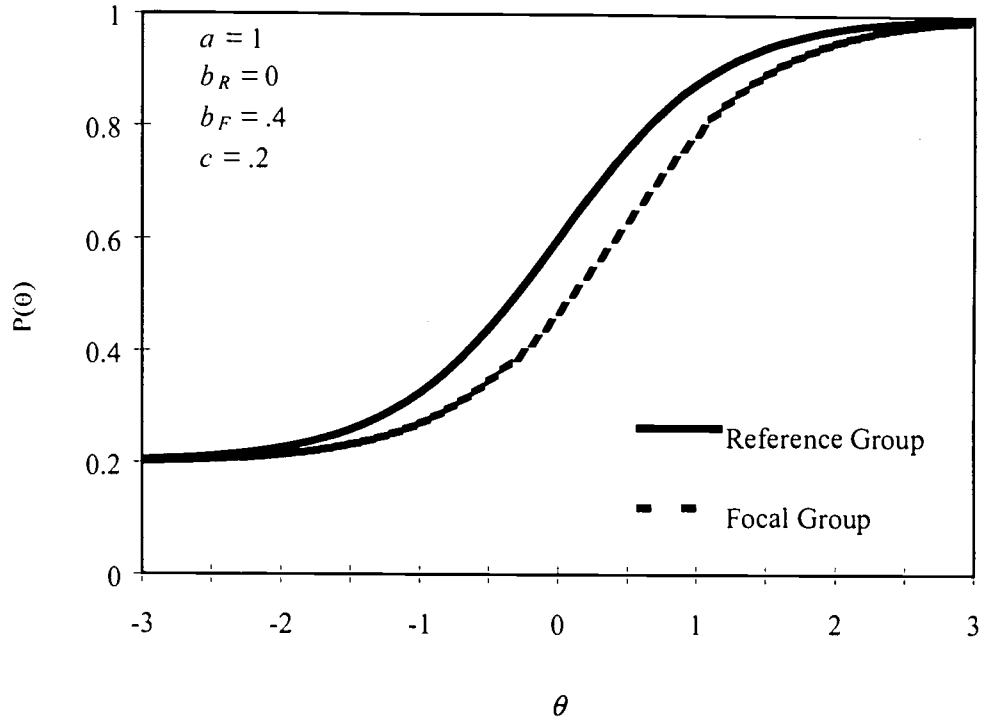

FIGURE 1. *Example of a DIF item used by Allen and Donoghue (1996).*

Allen and Donoghue (1996) simulated 150 replications of each of the 15 items (3 values of a times 5 values of b_R). They provided the mean and standard deviation of the $\hat{\Delta}$ values for the 150 replications for each item, which we used to create confidence intervals around $\hat{\Delta}$.

Results

Table 1 shows the mean $\hat{\Delta}$ values (in bold) from Allen and Donoghue (1996; their Table 4), as well as the Δ_{ZTL} values (Equation 6) and the $\Delta_{SM/RSP}$ values (Equation 8). The mean absolute difference between $\Delta_{SM/RSP}$ and $\hat{\Delta}$ is 0.11, whereas the mean absolute difference between Δ_{ZTL} and $\hat{\Delta}$ is 0.20, nearly double the amount for $\Delta_{SM/RSP}$. When the reference and focal groups have a large difference in their mean abilities (as in the simulated data in Allen & Donoghue, 1996), $\hat{\Delta}$ has a known estimation bias in the no-DIF case (e.g., Allen & Donoghue, 1996; Roussos & Stout, 1996). This bias can be roughly estimated by the values of mean $\hat{\Delta}$ in the "No DIF" column in Table 1 (all of these $\hat{\Delta}$ means would be within a standard error[2] or two of 0 if there were no bias). To correct for the bias, we subtracted these values from the mean $\hat{\Delta}$ values in the DIF column in Table 1 to obtain the column labeled "Corrected Mean $\hat{\Delta}$." We also calculated the standard errors for these differences (not shown in the table). Whereas only 4 of the 15 uncorrected mean $\hat{\Delta}$ values (the 4 with the smallest amount of

[2] Standard error is estimated by dividing the tabulated standard deviation *(SD)* of $\hat{\Delta}$ (provided in Table 1) by the square root of 150, the number of replications.

bias) are within two standard errors of the $\Delta_{SM/RSP}$ values, 13 of the 15 corrected values are within two standard errors of $\Delta_{SM/RSP}$. The bias correction reduced the mean absolute difference between corrected mean $\hat{\Delta}$ and $\Delta_{SM/RSP}$ to just 0.03. By contrast, Δ_{ZTL} is within two standard errors of mean $\hat{\Delta}$ for only one of the uncorrected values and for only one of the corrected values. Indeed, even after applying the bias correction, the mean absolute difference between Δ_{ZTL} and $\hat{\Delta}$ remained at 0.20. Even though Δ_{ZTL} exhibited a pattern similar to that of $\hat{\Delta}$, $\Delta_{SM/RSP}$ was more accurate in corresponding to the mean $\hat{\Delta}$ values.

TABLE 1

Comparison of MH DIF statistics from Allen and Donoghue (1996) with the Zwick, Thayer, and Lewis (1997) MH DIF parameter, Δ_{ZTL}; and the Spray and Miller (1992); Roussos, Schnipke, and Pashley (in press) MH DIF parameter, $\Delta_{SM/RSP}$.

		No DIF ($b_F = b_R$)		DIF Condition ($b_F = b_R + .4$)			DIF Condition Parameter Values	
						Corrected		
a	b_R	Mean $\hat{\Delta}$	*SD* of $\hat{\Delta}$	**Mean $\hat{\Delta}$**	*SD* of $\hat{\Delta}$	**Mean $\hat{\Delta}$**	Δ_{ZTL}	$\Delta_{SM/RSP}$
.5	-2	-.02	.22	**-.73**	.23	**-.71**	**-.76**	**-.73**
.5	-1	.00	.20	**-.66**	.19	**-.66**	**-.72**	**-.67**
.5	0	.04	.18	**-.52**	.19	**-.56**	**-.63**	**-.58**
.5	1	.06	.17	**-.36**	.18	**-.42**	**-.51**	**-.44**
.5	2	.08	.20	**-.18**	.20	**-.26**	**-.35**	**-.29**
1.0	-2	-.28	.33	**-1.71**	.28	**-1.43**	**-1.57**	**-1.49**
1.0	-1	-.17	.24	**-1.42**	.20	**-1.25**	**-1.47**	**-1.30**
1.0	0	-.02	.19	**-.96**	.18	**-.94**	**-1.19**	**-.95**
1.0	1	.07	.19	**-.43**	.22	**-.50**	**-.70**	**-.48**
1.0	2	.14	.19	**-.00**	.22	**-.14**	**-.26**	**-.15**
1.5	-2	-.54	.46	**-2.63**	.38	**-2.09**	**-2.38**	**-2.21**
1.5	-1	-.28	.23	**-2.07**	.22	**-1.79**	**-2.23**	**-1.85**
1.5	0	-.04	.21	**-1.17**	.21	**-1.13**	**-1.69**	**-1.17**
1.5	1	.09	.21	**-.31**	.22	**-.40**	**-.76**	**-.40**
1.5	2	.14	.24	**.05**	.23	**-.09**	**-.15**	**-.06**

Note. In the calculation of mean and standard deviation of $\hat{\Delta}$, Allen and Donoghue (1996) used 150 replications. The "No DIF" estimates from Allen and Donoghue indicate the amount of bias present in $\hat{\Delta}$ presumably caused by the large difference in mean proficiency between the reference and focal groups.

The results are even more dramatic when graphed. Figure 2 shows the absolute value of the corrected mean $\hat{\Delta}$'s from Allen and Donoghue (1996), $|\Delta_{ZTL}|$ calculated via Equation 6 based on the item parameters used to simulate the data, $|\Delta_{SM/RSP}|$ calculated via Equation 8, and $|\Delta_{2PL}|$ calculated via Equation 10. As shown in Figure 2, when a is .5 (upper left panel), there is not a very noticeable difference between $|\hat{\Delta}|$, $|\Delta_{ZTL}|$, and $|\Delta_{SM/RSP}|$, although $|\Delta_{2PL}|$ is clearly overestimating $|\hat{\Delta}|$ as item difficulty (b_R, and hence b_F) increases. When a is 1 (upper right panel), $|\hat{\Delta}|$ and $|\Delta_{SM/RSP}|$ are nearly indistinguishable, although $|\Delta_{ZTL}|$ overestimates $|\hat{\Delta}|$. Again, $|\Delta_{2PL}|$ greatly overestimates $|\hat{\Delta}|$ as item difficulty increases. When a is 1.5 (lower left panel), $|\hat{\Delta}|$ and $|\Delta_{SM/RSP}|$ are

again nearly indistinguishable, whereas $|\Delta_{ZTL}|$ and $|\Delta_{2PL}|$ overestimate $|\hat{\Delta}|$, even more than when $a = 1$. Although Δ_{ZTW} (Equation 5) is not shown, it is clear that *any* multiple of $(b_R - b_F)$ will not be an accurate summary of $\hat{\Delta}$ because $\hat{\Delta}$ is not a linear function of $(b_R - b_F)$ in the 3PL case (Roussos, Schnipke, & Pashley, in press).

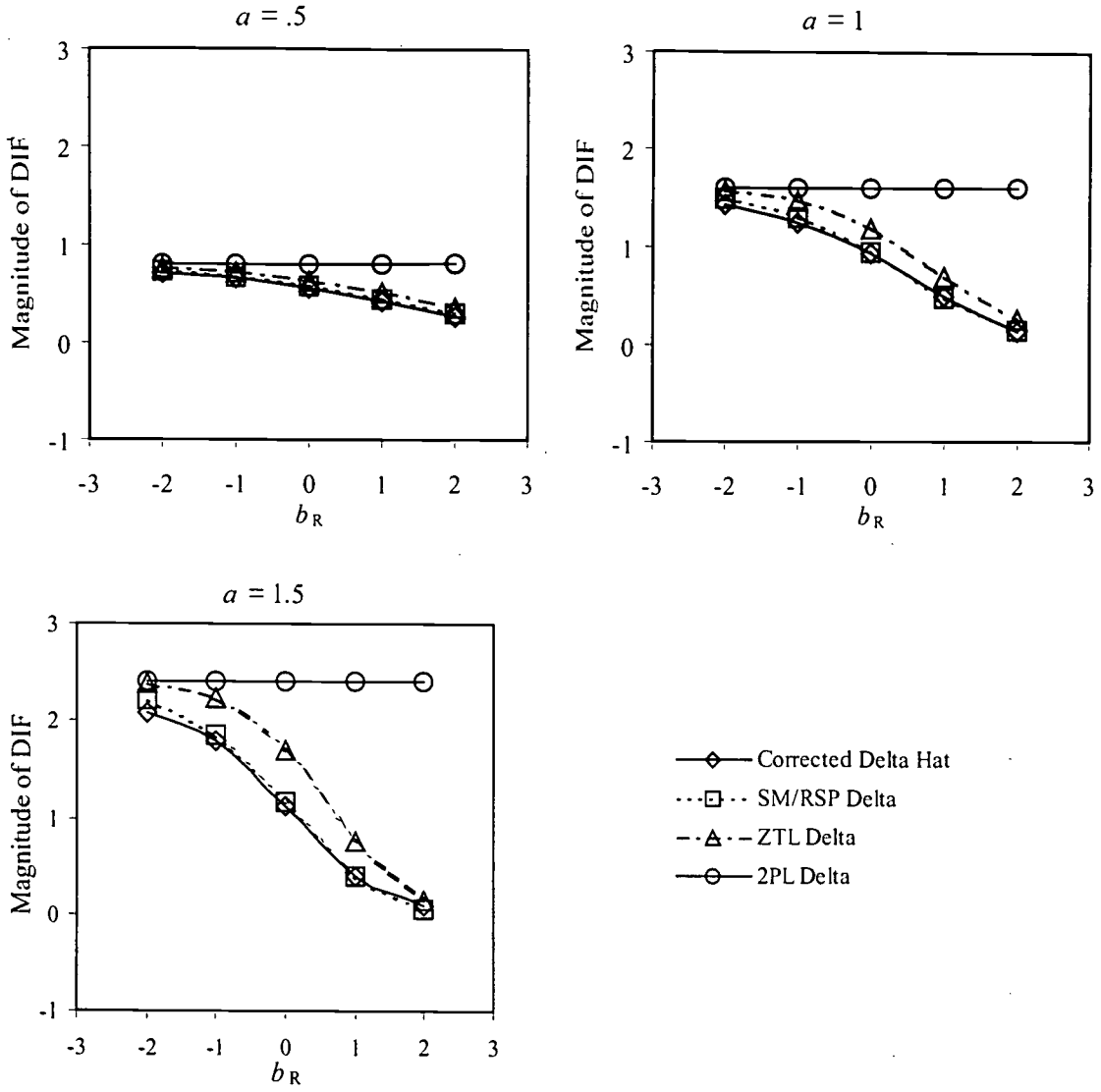

FIGURE 2. *Comparison of* $|\hat{\Delta}|$ *(from Allen & Donoghue, 1996) with various* $|\Delta|$ *parameters.*

The results suggest that for simulated DIF, a close correspondence exists between $\Delta_{SM/RSP}$ and $\hat{\Delta}$. Thus, inferences made about $\Delta_{SM/RSP}$ may be extended to the estimator ($\hat{\Delta}$). The Δ_{ZTL} parameter overestimates the value of $\hat{\Delta}$, especially as a increases. The inaccuracy in Δ_{ZTL} is not entirely surprising because the weights used by Δ_{ZTL} are not similar to the weights used by $\hat{\Delta}$.

Comparison of Δs With $\hat{\Delta}$ in Real Data

The results presented in the previous section were based on simulated data. Even though real item-response data on multiple-choice tests are often well approximated by the 3PL model, real data, unlike simulated data, do not perfectly correspond to the 3PL model. Thus, we may question whether realistic departures from the 3PL model might change the relationship between $\hat{\Delta}$ and $\Delta_{SM/RSP}$ or Δ_{ZTL}.

In this section, we artificially induce DIF in real data from a recent administration of the Law School Admission Test (LSAT). Using these real data, we compare how well $\hat{\Delta}$ corresponds to $\Delta_{SM/RSP}$, Δ_{ZTL}, and Δ_{2PL}.

Our investigation may be thought of as a test of the null hypothesis, H_0: $\hat{\Delta} - \Delta = 0$ (where Δ is either $\Delta_{SM/RSP}$, Δ_{ZTL}, or Δ_{2PL}). To give the most stringent test of this null hypothesis, we use large sample sizes (over 10,000 test takers in each group). This will result in the smallest possible standard error for $\hat{\Delta}$ so that even small differences between $\hat{\Delta}$ and Δ will be statistically significant. Values of Δ within two standard errors of $\hat{\Delta}$ will be considered as evidence for not rejecting the null hypothesis.

Method

To calculate $\Delta_{SM/RSP}$ and Δ_{ZTL}, we need to know the ability-distribution parameters and the studied-item parameters for both the reference and focal groups. The approach we chose was to first form artificial reference and focal groups by random assignment. Then we created a single "mock studied item" by using two real items, assigning one item to be the studied item for the reference group and the other item to be the same studied item for the focal group. We repeated this procedure numerous times to create a number of mock studied items. This procedure allowed us to (1) approximately match items on *a* and *c* (so that the only difference would be in item difficulty), (2) vary the difference in *b's* (to manipulate the amount of modeled DIF), and (3) vary the average *b* (to compare easy and hard items, for instance).

Specifically, based on previous dimensionality analyses (for example, Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996), an essentially unidimensional (Stout, 1987, 1990) subset of 49 items was selected from a recent form of the LSAT. Only item responses from white male test takers were selected in an attempt to create a homogeneous sample of test takers (to prevent potentially real DIF from interfering with our artificially induced DIF). Using BILOG (Mislevy & Bock, 1982) under the assumption of a standard normal ability distribution, 3PL item parameters were estimated for all 49 items using the responses from 20,092 white males. The white male test takers were then randomly split into two groups of 10,046 test takers each, with one group arbitrarily designated as the reference group and the other as the focal group. Finally, to create artificial studied items, each mock item was artificially formed by selecting one of the 49 items to be the studied item for the reference group and selecting some other item to be the studied item for the focal group. Thus, each mock studied item was in reality a pair of items with different items assigned to the reference and focal groups. The real responses (right/wrong) for these items were then used in the DIF analyses. Sixteen mock items were created in this manner.

Presented in Table 2 are the item parameter estimates and the observed proportion correct scores (based on all 20,092 white male test takers) for the 16 mock studied items. For convenience, the mock studied items are labeled from 1 to 16, and the item parameters are labeled with "R" and "F" subscripts to indicate which item parameters corresponded to the reference and focal groups. The observed proportion-correct scores

($\hat{p}_R$ and $\hat{p}_F$) confirm that items with higher b values were indeed more difficult (i.e., less likely to be answered correctly).

TABLE 2
Estimated item parameters and proportion-correct scores for the items used as the mock items

Item	b_R	b_F	c_R	c_F	a_R	a_F	$\hat{p}_R$	$\hat{p}_F$	$\lvert\hat{\Delta}\rvert$	$\lvert\Delta_{ZTL}\rvert$	$\lvert\Delta_{SM/RSP}\rvert$
Items with Negligible DIF											
1	0.75	0.76	.23	.25	.48	.46	.518	.532	0.27	0.14	0.15
2	-1.63	-1.72	.15	.13	.46	.46	.793	.799	0.13	0.09	0.08
3	0.66	0.76	.27	.25	.68	.46	.530	.532	0.07	0.10	0.01
4	0.30	0.21	.21	.22	.82	.81	.547	.568	0.31	0.27	0.27
5	1.03	1.13	.17	.21	.81	.94	.386	.376	0.13	0.02	0.00
Easy Items with Moderate to Large DIF											
6	-1.27	-0.71	.08	.08	.61	.64	.771	.680	1.31	1.22	1.24
7	-2.43	-1.49	.11	.12	.46	.44	.865	.764	1.79	1.76	1.70
8	-0.82	-1.63	.13	.12	.64	.65	.716	.835	1.85	1.98	1.93
9	-1.89	-0.94	.07	.07	.65	.62	.859	.715	2.48	2.52	2.43
10	-0.71	-1.80	.08	.09	.64	.65	.680	.850	2.72	2.79	2.74
Hard Items with Moderate to Large DIF											
11	1.42	1.03	.16	.17	.80	.81	.316	.386	0.75	0.72	0.77
12	1.43	1.92	.14	.16	.64	.75	.329	.269	0.70	0.72	0.75
13	1.92	1.42	.16	.16	.75	.80	.269	.316	0.62	0.54	0.64
14	1.13	1.92	.21	.16	.94	.75	.376	.269	1.17	1.43	1.39
15	1.03	1.92	.17	.16	.81	.75	.386	.269	1.38	1.41	1.39
16	0.81	1.92	.07	.16	.84	.75	.357	.269	0.97	1.17	1.03

Three forms of the MH DIF parameter, Δ_{2PL}, $\Delta_{SM/RSP}$, and Δ_{ZTL}, were calculated for each of the 16 mock studied items. The ability distributions of both the reference and focal groups were assumed to be standard normal distributions. Checking the $\hat{\theta}$ distributions of the two groups validated this assumption. The item parameter estimates from BILOG were treated as though they were the true item parameters in the calculation of the Δ's. The proportion of test takers from the reference group and focal group was, by design, .5 (equal proportions from each). The Δ_{2PL} values were calculated using the average estimated a parameter of the mock item, which can be calculated from Table 2.

The MH DIF statistic, $\hat{\Delta}$, was calculated for each of the 16 mock items. The matching criterion was the score on the studied item plus the score on a subset of the remaining 47 items. To ensure a matching criterion that was as unidimensional as possible, 10 items suspected of speededness[3] were excluded from being used in the matching criterion, although some of these items were used as mock studied items. Thus, there were 38 to 40 items on the matching criterion, depending on whether neither, one, or both of the two items used to form the mock item were one of the 10 items excluded from the matching criterion.

[3] The 10 items that were excluded due to possible speededness were the five items at the end of each of the two sections where the time limit may affect performance on the items.

Results and Discussion

Figure 3 presents $|\Delta_{2PL}|$, $|\Delta_{SM/RSP}|$, $|\Delta_{ZTL}|$, and $|\hat{\Delta}|$ for each mock item. The first five mock items were chosen so that the difference in the difficulty parameters, $|b_R - b_F|$, was small and, thus, would result in negligible DIF. As shown in Figure 3, $|\Delta_{2PL}|$, $|\Delta_{SM/RSP}|$, $|\Delta_{ZTL}|$, and $|\hat{\Delta}|$ were all small (less than 0.5 in all cases). Notice that the correspondence between $|\Delta_{SM/RSP}|$ and $|\hat{\Delta}|$ and between $|\Delta_{ZTL}|$ and $|\hat{\Delta}|$ was closer than that between $|\Delta_{2PL}|$ and $|\hat{\Delta}|$. For all five of these negligible-DIF mock items, $\Delta_{SM/RSP}$ and Δ_{ZTL} were within two standard errors of $\hat{\Delta}$, whereas Δ_{2PL} was within two standard errors for only two of the five items (items 2 and 4).

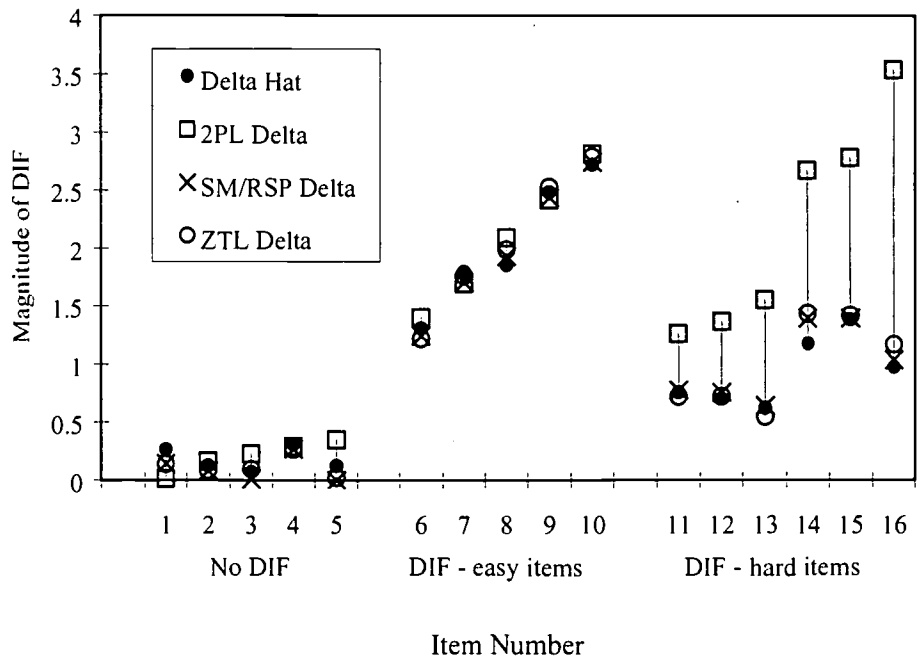

FIGURE 3. *Comparison of* $\hat{\Delta}$, Δ_{2PL}, $\Delta_{SM/RSP}$, *and* Δ_{ZTL} *in "real" data.*

The second set of five mock items was chosen to contain relatively easy items (moderate to large negative b parameters). As shown in Figure 3, for items 6-10, $|\Delta_{SM/RSP}|$, $|\Delta_{ZTL}|$, and $|\hat{\Delta}|$ were approximately equal to each other. For all five of these mock items, $\Delta_{SM/RSP}$ and Δ_{ZTL} were within two standard errors of $\hat{\Delta}$, while Δ_{2PL} was within two standard errors of $\hat{\Delta}$ for four of the five items (items 6, 7, 9, and 10).

Finally, the last six mock items were chosen to be relatively difficult items (moderate to large positive b parameters). Figure 3 shows that $\Delta_{SM/RSP}$ and Δ_{ZTL} both corresponded well to $\hat{\Delta}$, although Δ_{2PL} substantially overpredicts $\hat{\Delta}$. For five of these six mock items, $\Delta_{SM/RSP}$ was within two standard errors of $\hat{\Delta}$ (items 11, 12, 13, 15, and 16), whereas Δ_{ZTL} was within two standard errors of $\hat{\Delta}$ for four of the six items (items 11, 12, 13, and 15). However, Δ_{2PL} was within two standard errors of $\hat{\Delta}$ for *none* of the six items. The 2PL formula, $4a(b_R - b_F)$, does not work well with moderately to highly difficult items in the 3PL case.

For these real data, $\Delta_{SM/RSP}$ and Δ_{ZTL} performed similarly well. These items tended to have low values of a (see Table 2), however, and the differences between $\Delta_{SM/RSP}$ and Δ_{ZTL} were minimal for small values of a in the

simulated data. Differences between $\Delta_{SM/RSP}$ and Δ_{ZTL} were noted primarily for large values of a in the simulated data, and in that case $\Delta_{SM/RSP}$ corresponded to $\hat{\Delta}$ better than Δ_{ZTL} did. Thus we would expect in real data that $\Delta_{SM/RSP}$ would correspond to $\hat{\Delta}$ better than Δ_{ZTL} would as a increases. This is an important issue to be addressed by future work with mock items.

Conclusions

Three MH DIF parameters (Δ_{2PL}, $\Delta_{SM/RSP}$, and Δ_{ZTL}) were compared with values of the MH DIF statistic ($\hat{\Delta}$) in simulated and real data. In both simulated and real data, Δ_{2PL} overestimated $\hat{\Delta}$, especially as item difficulty increased. In both simulated and real data, $\Delta_{SM/RSP}$ was nearly identical to $\hat{\Delta}$ in all conditions. In simulated data, Δ_{ZTL} slightly overestimated $\hat{\Delta}$, especially as item difficulty and discrimination increased, although it performed well in real data. The real items used in the present study tended to have low discrimination values, and based on results with simulated items, we would expect Δ_{ZTL} to perform less well for higher values of item discrimination. Thus $\Delta_{SM/RSP}$ appears to be the best parameter for the MH DIF statistic overall.

References

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

Donoghue, J., 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: Erlbaum.

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

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

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

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

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type 1 error performance. *Journal of Educational Measurement, 33*, 215-230.

Shealy, R., & 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*, 159-194.

Spray, J. A., & Miller, T. R. (1992). *Performance of the Mantel-Haenszel statistic and the standardized difference in proportion correct when population ability distributions are incongruent* (Research Report No. 92-1). Iowa City, IA: American College Testing.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Stout, W. F., Habing, B., Douglas, J. A., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331-354.

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: Erlbaum.

Zwick, R., Thayer, D. T., & Lewis, C. (1997). *An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis* (Research Report No. RR-97-21). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18*, 121-140.

Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32*, 341-363.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").