Listening. Learning. Leading.®

Research Report
ETS RR-13-16

An Investigation of the Efficacy of Criterion Refinement Procedures in Mantel-Haenszel DIF Analysis

Rebecca Zwick

Lei Ye

Steven Isham

September 2013

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

An Investigation of the Efficacy of Criterion Refinement Procedures in Mantel-Haenszel DIF Analysis

Rebecca Zwick, Lei Ye, and Steven Isham
Educational Testing Service, Princeton, New Jersey

September 2013

Action Editor: Marna Golub-Smith

Reviewers: Longjuan Liang and Tim Moses

Abstract

Differential item functioning (DIF) analysis is a key component in the evaluation of the fairness and validity of educational tests. Although it is often assumed that refinement of the matching criterion always provides more accurate DIF results, the actual situation proves to be more complex. To explore the effectiveness of refinement, we conducted a simulation study consisting of 40 conditions that varied in terms of amount and pattern of DIF, sample sizes, and ability distributions. We found that the effectiveness of refinement was heavily dependent on whether DIF was balanced (with positive DIF values compensating for negative DIF values) or unbalanced (all in one direction). In balanced conditions, the unrefined method generally produced better results, whereas, in unbalanced conditions, the opposite was true. In the absence of information about the pattern of DIF, it is probably best to choose the refined method because it is only slightly disadvantageous in balanced conditions, whereas the unrefined method can be substantially disadvantageous in certain unbalanced conditions.

Key words: differential item functioning (DIF), item bias, refinement, purification, Mantel-Haenszel

Acknowledgments

We would like to thank Marna Golub-Smith, Shelby Haberman, Longjuan Liang, and Tim Moses for their comments on an earlier version of this paper. We also appreciate the data analysis assistance we received from Jonathan Guglielmon.

Differential item functioning (DIF) analysis is a key component in the evaluation of the fairness and validity of educational tests. As part of their standard operations, testing companies conduct DIF analyses on thousands of items per year. In an attempt to improve the accuracy of DIF assessment, many testing programs make use of criterion refinement. Refinement is intended to improve the quality of the matching criterion by removing items identified as having DIF in a preliminary round of analysis and then performing the DIF analysis again. Although it is often assumed that refinement always provides superior results, the actual situation proves to be more complex.

In an informal report (Lord, 1976) that later appeared as a book chapter (Lord, 1977), Frederic Lord made what is perhaps the first published reference to criterion refinement (though he did not use either that term or *purification*, a term used in much of the early literature in this area). Lord incorporated a refinement procedure, an idea he later attributed to Gary Marco (Lord, 1980, p. 220), as part of an item response theory–based study of item bias on the *SAT*® exam. The recommendation to use criterion refinement when applying the Mantel-Haenszel (MH; Mantel & Haenszel, 1959) DIF procedure was made by Holland and Thayer (1986a, 1986b, 1988, p. 42), who stated that the recommendation was based on a conjecture. They cited a similar suggestion made by Kok, Mellenbergh, and van der Flier (1985) in connection with a logit-based DIF procedure. The recommendation to use refinement appeared in Educational Testing Service (ETS) DIF policy memos in the late 1980s and was repeated by Dorans and Holland (1993, pp. 60–61) in their review chapter on DIF detection.

Some recent findings, however, did not support the use of refinement. In the course of a larger study, Zwick, Ye, and Isham (2012) compared refined and unrefined MH results for some simulated item response data and found a slight advantage for the unrefined results. This finding was in contrast to much of the existing literature. For example, Clauser, Mazor, and Hambleton (1993) conducted what appears to be the first simulation study of refinement and concluded that refined results were "equal or superior" (p. 269) to unrefined results, both in terms of Type I and Type II error. Recent reviews of the refinement literature (Colvin & Randall, 2011; French & Maller, 2007) concluded that refinement was typically found to have a favorable effect on the accuracy of DIF procedures.

Although the initial Zwick et al. (2012) refinement analyses were in some ways similar to those of Clauser et al. (1993), the two studies differed in terms of the pattern of the DIF that was

modeled. In the Zwick et al. (2012) simulation, the true DIF values had an average near zero across the 34 items in the test; that is, in a rough sense, positive and negative DIF were balanced. In the Clauser et al. study, all DIF was in one direction—against the focal group. We conjectured that if DIF is balanced, the contaminated matching criterion that is used in an unrefined analysis might, nevertheless, be an adequate measure of proficiency. (Wang & Su, 2004, made a similar speculation.) Applying refinement may serve mainly to reduce the precision of the matching criterion, degrading the results. In the Clauser et al. study, however, the unrefined matching criterion had a systematic bias against focal group members that was reduced by refinement. The disparity in results between the Clauser et al. study and the Zwick et al. (2012) analysis prompted us to carry out a comprehensive simulation study comparing refined and unrefined DIF results.

Unlike previous simulation studies of refinement, our study examined the accuracy of a DIF flagging rule, used at ETS, that involves both effect size and statistical significance, as well as a rule based solely on the MH chi-square, as in Clauser et al. (1993). Also, in evaluating the simulation outcomes, we examined the properties of the unrefined and refined MH estimates (variance, bias, root mean square residual [RMSR]) in addition to the Type I rate and power associated with the unrefined and refined flagging procedures.

Method

Data Simulation

Our main simulation consisted of 40 conditions that varied in terms of the following factors:

- Percentage of items on the test with DIF (0%, 10%, or 20%). Items without DIF had identical item response functions for the reference and focal groups.
- Pattern of DIF. Balanced DIF (i.e., DIF in both directions, constructed so that the sum of the true DIF values was approximately 0) or unbalanced DIF (all DIF in one direction).
- Focal group distribution. The focal group ability distribution was either standard normal (N(0,1)) or normal with a mean of -1 and a variance of 1 (N(-1,1)). The reference group distribution was always N(0,1).
- Length of test (20 or 80 items).
- Reference and focal group sample sizes ($n_R = n_F = 500$ or $n_R = 200$, $n_F = 50$).

The second sample-size condition, which involves samples smaller than those used in typical DIF analyses, was included to determine whether refinement methods would operate

differently when sample sizes dropped below conventional levels. As described later, this proved to be the case.

Differences between refined and unrefined methods in Type I error and detection rates were very small in the 80-item tests, even when 20% of the items had DIF, perhaps because the number of non-DIF items (always at least 64) was sufficient to allow for reasonably accurate matching. Because the refined and unrefined analysis methods performed almost identically for these tests, only the results of the 20-item analyses are included here. The 20 simulation conditions that involved 20-item tests are summarized in the first four columns of Table 1.

Table 1

Summary of Simulation Conditions for 20-Item Tests

Condition	Focal distribution[a] and group sample sizes	% of items with DIF	DIF pattern	Average true DIF over 20 items	Average |true DIF| over 20 items	Average true DIF for DIF items only	Average |true DIF| for DIF items only
1A	N(0,1) $n_R = n_F = 500$	0	No DIF	0	0	0	0
1B	″	10	Balanced	-.00065	.163	-.007	1.626
1C	″	20	″	-.0008	.338	-.004	1.689
1D	″	10	Unbalanced	-.169	.169	-1.693	1.693
1E	″	20	″	-.334	.334	-1.672	1.672
2A	N(0,1); $n_R = 200$, $n_F = 50$	0	No DIF	0	0	0	0
2B	″	10	Balanced	-.00065	.163	-.007	1.626
2C	″	20	″	-.0008	.338	-.004	1.689
2D	″	10	Unbalanced	-.169	.169	-1.693	1.693
2E	″	20	″	-.334	.334	-1.672	1.672
3A	N(-1,1) $n_R = n_F = 500$	0	No DIF	0	0	0	0
3B	″	10	Balanced	-.00065	.163	-.007	1.626
3C	″	20	″	-.0008	.338	-.004	1.689
3D	″	10	Unbalanced	-.169	.169	-1.693	1.693
3E	″	20	″	-.334	.334	-1.672	1.672
4A	N(-1,1); $n_R = 200$, $n_F = 50$	0	No DIF	0	0	0	0
4B	″	10	Balanced	-.00065	.163	-.007	1.626
4C	″	20	″	-.0008	.338	-.004	1.689
4D	″	10	Unbalanced	-.169	.169	-1.693	1.693
4E	″	20	″	-.334	.334	-1.672	1.672

Note. DIF = differential item functioning.

[a] The reference group distribution was always N(0,1). True DIF values for individual items are given in Table 2.

Item responses were generated using the three-parameter logistic (3PL) model, with 500 replications per item per condition. As a starting point, we used the item parameters from Clauser et al. (1993, p. 272) as reference group item parameters for our 80-item conditions. These parameters were originally "taken from parameter estimates of actual test items … of the Graduate Management Admissions Test" (Clauser et al, p. 271). For our 20-item conditions, we used a subset of the 80 items.

To induce DIF in the selected items, we added or subtracted 0.6 from the reference group difficulty parameter to obtain the focal group difficulty parameter. All discrimination and difficulty parameters used in our 20-item simulations are listed in Table 2. Guessing parameters for all items were set to .20, as in Clauser et al. Using the formulation in Zwick, Thayer, and Lewis (2000, p. 234) for translating item parameters to the MH metric, the true DIF, expressed in the MH metric, ranged from 0 to 1.754 in magnitude across the 20 conditions in Table 1. The true DIF values corresponding to each focal group difficulty parameter are shown in Table 2. Summary statistics for the true DIF values for each condition are shown in Columns 5–8 in Table 1. As shown in the rightmost column, the amount of DIF for the balanced and unbalanced conditions is roughly equivalent in terms of the absolute magnitude of the true MH values.

The ETS DIF classification system is based on the size and statistical significance of the MH delta difference statistic, *MH D-DIF*, defined as follows:

$$MH\ D\text{-}DIF = -2.35(\ln(\hat{\beta}_{MH})), \quad (1)$$

where $\hat{\beta}_{MH}$ is the MH odds ratio estimate (Mantel & Haenszel, 1959). If the MH chi-square test (with continuity correction; see Holland & Thayer, 1988) is not significant at $\alpha = .05$ or if |*MH D-DIF*| is less than 1, the item is considered to be an *A*, or negligible-DIF, item. Items with large DIF are referred to as *C* items. A *C* identification requires that |*MH D-DIF*| be greater than 1.5 and be statistically different from 1 at $\alpha = .05$. As shown by Holland (see Zwick, 2012), the statistical test involves determining whether the quantity (|*MH D-DIF*|-1)/*SE*(*MH D-DIF*) exceeds 1.645, where *SE*(*MH D-DIF*) is the estimated standard error of *MH D-DIF* (see Holland & Thayer, 1988). Items that are neither *A* nor *C* items are considered *B*, or intermediate-DIF, items.

Table 2

Item Parameters for the Simulation Study

Item	Discrimination	Reference group item difficulty	Focal group item difficulty and true DIF: Conditions 1B, 2B, 3B, 4B	Focal group item difficulty and true DIF: Conditions 1C, 2C, 3C, 4C	Focal group item difficulty and true DIF: Conditions 1D, 2D, 3D, 4D	Focal group item difficulty and true DIF: Conditions 1E, 2E, 3E, 4E
1	0.75	-1.97				-1.37 (-1.718)
2	0.73	-1.60	-1.00 (-1.632)	-1.00 (-1.632)	-1.00 (-1.632)	-1.00 (-1.632)
3	0.64	-1.55				
4	0.81	-0.62		-1.22 (1.751)		
5	0.39	0.04				
6	0.85	-0.37				0.23 (-1.584)
7	0.87	-0.75		-0.15 (-1.754)	-0.15 (-1.754)	-0.15 (-1.754)
8	0.78	-0.05				
9	0.45	-1.49				
10	0.61	-0.53				
11	0.98	0.31				
12	0.50	0.80				
13	0.29	-1.00				
14	0.70	1.05				
15	1.02	0.64	0.04 (1.619)	0.04 (1.619)		
16	1.16	1.11				
17	0.48	2.12				
18	0.65	1.19				
19	1.01	0.91				
20	0.53	0.87				

Note. True DIF values corresponding to each focal group difficulty parameter are shown in parentheses in Columns 3–6. Negative true DIF values indicated DIF against the focal group, while positive values indicate DIF against the reference group. Except where indicated otherwise, focal group difficulty parameters were the same as reference group parameters. The reference group discrimination and difficulty parameters in Columns 2 and 3 are the same as those for items 41–60 in Clauser et al. (1993). All guessing parameters were set to .20. The 3PL model included a scale factor of 1.7. DIF = differential item functioning.

Consistent with these criteria, items with true DIF of at least 1.5 in magnitude in our simulation were considered true *C* items. This includes all DIF items in the 20-item simulations (see Table 2).

DIF analyses were conducted with and without refinement. Our refinement procedure was identical to that used operationally at ETS: An initial DIF run is conducted to identify *C* items. The matching criterion is the total score, including the studied item. In a second run, these

C items are deleted from the matching criterion. An exception to this is that the studied item itself is always included in the matching criterion.

We present three different types of results: an assessment of the properties of the *MH D-DIF* statistic itself; an assessment of detection and Type I error for the ETS *C* rule, described above; and an assessment of the DIF detection and Type I error rates for a flagging rule based only on the MH chi-square test.

Properties of the *MH D-DIF* Statistic

For the refined and unrefined *MH D-DIF* statistic for each item in each condition, we obtained the signed bias $B(\hat{\omega})$, squared bias $B^2(\hat{\omega})$, variance $Var(\hat{\omega})$, and root mean square residual $RMSR(\hat{\omega})$ of the *MH D-DIF* statistics under balanced and unbalanced patterns of DIF for the conditions described above. These quantities are defined as follows for each item (with item subscripts omitted for simplicity):

$$B(\hat{\omega}) = \bar{\hat{\omega}} - \omega, \tag{2}$$

$$B^2(\hat{\omega}) = (\bar{\hat{\omega}} - \omega)^2, \tag{3}$$

$$Var(\hat{\omega}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\omega}_r - \bar{\hat{\omega}})^2}, \tag{4}$$

and

$$RMSR(\hat{\omega}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\omega}_r - \omega)^2} = \sqrt{B^2(\hat{\omega}) + Var(\hat{\omega})}, \tag{5}$$

where $\hat{\omega}_r$ represents the *MH D-DIF* value for replication *r*, $\bar{\hat{\omega}} = \frac{1}{R}\sum_{i=1}^{R}\hat{\omega}_r$ is the average of $\hat{\omega}_r$ across replications, ω is the true DIF value, and R = 500 is the number of replications.

Type I Error Rates and Detection Rates for the ETS *C* Rule

For the refined and unrefined *MH D-DIF* statistic for each item in each condition, we computed the flagging rate as the number of replications for which the item was identified as a *C* item divided by 500, the number of replications. In the case of non-DIF items, the flagging rate is

the Type I error rate; in the case of DIF items, it is the detection rate. We computed Type I error rates for all the non-DIF items in the 16 conditions that contained some DIF items, as well as in the four conditions that contained no DIF items (1A, 2A, 3A, and 4A).

Type I Error Rates and Detection Rates for a Flagging Rule Based Only on MH Chi-Square

The ETS DIF classification rules, which are based on both effect size and statistical significance, were devised to minimize the role of sample size in DIF classification. The use of a DIF flagging rule based only on a statistical significance test is not recommended because results will be heavily influenced by sample size. An *MH D-DIF* value that would appear to represent extreme DIF in one test administration might not lead to statistically significant results in a smaller administration. Also, a tiny *MH D-DIF* value would lead to flagging if the sample was sufficiently large. Despite the disadvantages of relying solely on significance tests, we have included some results for a flagging rule based only on the MH chi-square statistic because the flagging rates are, in some respects, easier to interpret. The ETS *C* rule, which requires that two separate criteria be met—an effect size criterion and a statistical significance criterion—cannot be said to have a nominal Type I error rate in the usual sense (although obviously, the Type I error rate is expected to be below the alpha level for the associated significance test). The MH chi-square tests, by contrast, were conducted at a nominal alpha level of .05, facilitating the interpretation of the Type I error results and the comparison of refined and unrefined methods.

In the refined version of the chi-square procedure, items with a significant chi-square in a preliminary analysis are deleted from the matching criterion in the final analysis. As in the ETS *C* rule analysis, an exception to this rule is that the studied item is always included in the matching criterion.

Results

The results concerning the properties of DIF estimators are discussed first, followed by our findings on Type I error rates and DIF detection rates.

Properties of DIF Estimators

Tables 3 and 4 show average signed bias, average squared bias, average variance, and average RMSR for $n_R = n_F = 500$ and $n_R = 200$, $n_F = 50$, respectively, for the case of 20% DIF.

Patterns were similar but less prominent for the 10% conditions (not shown). The averages in Tables 3 and 4 are taken over the four items that had DIF in each condition (not over all 20 items in the condition).

Table 3

Average Signed Bias, Squared Bias, Variance, and RMSR of MH D-DIF Statistics for DIF Items Under 20% DIF Conditions, $n_R = n_F = 500$

Condition	Method	Average signed bias	Average squared bias	Average variance	Average RMSR
Balanced Condition 1C	Refined	-.006	.005	.158	.402
	Unrefined	-.012	.000	.154	.390
	Average	-.009	.003	.156	.396
Unbalanced Condition 1E	Refined	.306	.094	.187	.527
	Unrefined	.368	.136	.181	.560
	Average	.337	.115	.184	.543
Balanced Condition 3C	Refined	-.419	.212	.147	.585
	Unrefined	-.288	.099	.150	.494
	Average	-.353	.155	.149	.539
Unbalanced Condition 3E	Refined	.038	.023	.179	.448
	Unrefined	.174	.050	.172	.470
	Average	.106	.037	.175	.459

Note. Averages are based on the four DIF items in each condition. DIF = differential item functioning, n_F = focal group, n_R = reference group, RMSR = root mean square residual.

Table 4

Average Signed Bias, Squared Bias, Variance, and RMSR of MH D-DIF Statistics for DIF Items Under 20% DIF Conditions, $n_R = 200$, $n_F = 50$

Condition	Method	Average signed bias	Average squared bias	Average variance	Average RMSR
Balanced Condition 2C	Refined	.058	.005	1.153	1.070
	Unrefined	.060	.007	1.144	1.066
	Average	.059	.006	1.149	1.068
Unbalanced Condition 2E	Refined	.376	.151	1.240	1.165
	Unrefined	.411	.180	1.239	1.176
	Average	.393	.165	1.239	1.171
Balanced Condition 4C	Refined	-.289	.105	.969	1.034
	Unrefined	-.245	.075	.948	1.010
	Average	-.267	.090	.959	1.022
Unbalanced Condition 4E	Refined	.173	.047	1.032	1.038
	Unrefined	.224	.066	1.027	1.044
	Average	.199	.057	1.029	1.041

Note. DIF = differential item functioning, n_F = focal group, n_R = reference group, RMSR =root mean square residual.

For the average variances and average RMSRs, the most obvious pattern is that they were, of course, much larger in the small-sample case (Table 4) than in the large-sample case (Table 3). Overall, the refined and unrefined DIF statistics performed quite similarly for the small-sample conditions. The average RMSRs were roughly equal to 1, meaning that this was the average disparity between *MH D-DIF* values and their underlying parameters under these conditions. In the large-sample conditions, average RMSRs ranged from .39 to .59. Average RMSRs were always smaller for the unrefined methods in the balanced cases and smaller for the refined methods in the unbalanced cases, though differences were slight in the small-sample conditions. Average variances for refined and unrefined methods were similar in both sample size conditions.

Patterns of bias in *MH D-DIF* statistics. Regarding the bias of the DIF statistics, several patterns were apparent:

1. For the balanced conditions involving the N(0,1) focal group (1C and 2C), biases were quite small for both the refined and unrefined DIF methods.
2. For the balanced conditions involving the N(-1,1) focal group (3C and 4C), biases were large and negative. The bias for the refined DIF method was larger than for the unrefined. This was especially true in the large-sample Condition 3C (Table 3), where the bias for refined was -.42, compared to -.29 for unrefined.
3. For unbalanced conditions, biases were positive. They were larger in conditions involving the N(0,1) focal group (1E and 2E) than in conditions involving the N(-1,1) focal group (3E and 4E). Biases in the unbalanced conditions were larger for the unrefined than for the refined methods. The discrepancy was particularly notable in the large-sample Condition 3E (Table 3), where the bias for unrefined was .17, compared to .04 for refined.

The observed patterns of bias are likely to result from the combined influences of three separate effects: (a) apart from any considerations of DIF, a difference in the average ability for the two groups will produce systematic errors in matching, unless the matching criterion is very reliable; (b) unbalanced negative DIF will (other things being equal) lead to systematic matching errors in the opposite direction; and (c) the self-norming nature of the MH statistic, discussed below, can result in bias in unbalanced DIF conditions. Each of these effects is now considered in more detail.

In the conditions in which the reference group and focal group have different ability distributions, systematic matching errors are to be expected when the matching criterion is not perfectly reliable. (In our 20-item conditions, the reliability, averaged over four representative simulation conditions, each with 500 replications, was .71.) Using a classical test theory framework, we can express the expected true score *T*, given a particular observed matching criterion score *x* for a member of Group *G* as follows (see Lord & Novick, 1968, p. 65):

$$E_G(T|X = x) = \rho x + (1 - \rho)\mu_{xG}, \quad (6)$$

where ρ is the test reliability and μ_{xG} is the mean test score for Group G. For now, let us assume that *X* itself is not distorted by DIF. For a given matching score *x*, the lower scoring group will tend to have lower ability (*T*) than the "matched" members of the higher scoring group because it has a lower mean μ_{xG}. This will lead to the appearance of DIF even where none exists (see Zwick, 1990) and will tend to make any existing DIF against the focal group (i.e., negative DIF) appear to be more severe. (DIF against the reference group will appear less severe, as discussed below.) This phenomenon is consistent with the negative bias in balanced Conditions 3C and 4C, where the matching variable itself is not systematically distorted by DIF.

In the unbalanced conditions, there is a second force at work. Here, all the DIF is against the focal group, so that focal group members are likely to be assigned values on the matching criterion *X* that are too low, relative to their ability *T*. Hence, in situations where the ability groups for the reference and focal groups are the same (1E and 2E), focal group members will tend to perform better *relative to the reference group* than they would if the groups were matched on *T*, given their (downwardly biased) scores on the matching criterion, *X*. This produces positive bias in the *MH D-DIF* values, making it harder to detect the DIF, all of which is negative in these conditions. This distortion is made less severe through the use of refinement.

In the unbalanced conditions with N(-1,1) focal groups (3E and 4E), both these forces are at work, in opposite directions. Biases are notably less than in unbalanced conditions 1E and 2E.

The third factor influencing the bias stems from the disparity between the amount of DIF in the matching criterion in a given analysis and the average MH value. Mantel-Haenszel DIF statistics have a self-norming property: When total score is the matching criterion, the *MH D-DIF* statistics for the items contributing to the total score are negatively correlated. When there are two items, it is easy to show that the two *MH D-DIF* statistics are constrained to have an

average value of exactly zero; hence, their correlation is -1. Demonstrating the precise nature of the constraints when there are more than two items has so far proved elusive, but results show that the average of the (unrefined) *MH D-DIF* values tends to be close to zero when total score is the matching criterion. This result holds regardless of the underlying pattern of DIF. For example, in Condition 1E of our study, for which the average true DIF was -.334 (see Table 1), the median of the average MH values across the 500 replications was -.06 for the unrefined analysis. The farther the average true DIF from zero, the more biased the MH statistics.[1]

Type I Error Rates for the ETS *C* Rule

The Type I error rates for the refined and unrefined procedures for all conditions were very low—below 1%. (This is quite typical for the ETS *C* rule; see Zwick, 2012.) For the four no-DIF conditions, average Type I error rates are shown in Table 5. For the items without DIF in the remaining conditions, Type I error rates are given in the DIF absent columns of Tables 6 through 9.

The only factor that had a substantial effect on Type I error was sample size: The average Type I error for $n_R = n_F = 500$ was .034, while for $n_R = 200$, $n_F = 50$, it was .687. (Note that these averages are in the percent, rather than the proportion metric.) The average Type I error rate for non-DIF items in unbalanced conditions was slightly higher than the average error rate in balanced or no-DIF conditions. In balanced conditions, refined DIF analyses tended to yield higher Type I error rates than unrefined analyses, while in unbalanced conditions, unrefined methods tended to have higher rates than refined methods. However, differences between refined and unrefined methods were small.

DIF Detection Rates for the ETS *C* Rule

The DIF present columns of Tables 6 through 9 show detection rates for the 16 conditions that contain DIF items. Table 6 shows the average DIF detection rate for the 20% DIF conditions with $n_R = n_F = 500$. The table shows that in the balanced conditions, the unrefined method has a higher detection rate (53.9% vs. 47.4% in Condition 1C and 44.4% vs. 42.1% in 3C), while in the unbalanced conditions, the refined method has a higher detection rate (22% vs. 18.0% in 1E and 45.1% and 32.9% in 3E). The advantage of the unrefined method in the balanced conditions is greater in the N(0,1) focal group (Condition 1C) than in the N(-1,1) focal group (3C), whereas

the advantage of the refined method in the unbalanced conditions is greater in the N(-1,1) focal group (3E) than in the N(0,1) focal group (1E).

Table 5

DIF Flagging Rates (in Percent Metric) for ETS C Rule Under Conditions Without DIF

Sample size	Condition	Method	Flagging rate
$n_R = n_F = 500$	1A	Refined	0.010
		Unrefined	0.010
	3A	Refined	0.040
		Unrefined	0.040
$n_R = 200$, $n_F = 50$	2A	Refined	0.540
		Unrefined	0.570
	4A	Refined	0.780
		Unrefined	0.830

Note. Each entry is an average over 20 items, with 500 replications per item. Standard errors of table entries range from 0.01 to 0.09. DIF = differential item functioning, $n_{F\,=}$ focal group, n_R = reference group.

Table 6

DIF Flagging Rates for ETS C Rule Under 20% DIF Conditions, $n_R = n_F = 500$

Method	N(0, 1) focal group				N(-1, 1) focal group			
	Balanced: Condition 1C		Unbalanced: Condition 1E		Balanced: Condition 3C		Unbalanced: Condition 3E	
	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present
Refined	0.000	47.4	0.025	22.0	0.125	42.1	0.075	45.1
Unrefined	0.000	53.9	0.025	18.0	0.037	44.4	0.100	32.9
Average	0.000	50.7	0.025	20.0	0.081	43.3	0.088	39.0

Note. Entries in DIF absent columns are averages over 16 items. Entries in DIF present columns are averages over four items. For entries in the refined and unrefined rows, standard errors range from 0.00 to 1.12. For the average row, standard errors range from 0.00 to 0.79. DIF = differential item functioning.

Tables 7, 8, and 9 give the detection rates for the remaining conditions. Table 7 pertains to conditions in which $n_R = n_F = 500$, as in Table 6. However, the concentration of DIF is lower: Only two of the 20 items have DIF. These two items are a subset of the four DIF items included in Table 6. Table 7 shows a pattern very similar to that of Table 6: In the balanced conditions, the

unrefined method has a higher detection rate, while in the unbalanced conditions, the refined method has a higher detection rate. Tables 8 and 9, which pertain to the 20% and 10% DIF conditions, respectively, for the small-sample-size condition (n_R = 200, n_F = 50), show that, as noted earlier, the refined and unrefined approaches perform very similarly in these circumstances. In both the balanced and unbalanced cases, the method with the higher detection rate was most often the one with the lower average bias (Tables 3 and 4) and, somewhat surprisingly, the *lower* Type I error rate.

Table 7

DIF Flagging Rates for ETS C Rule Under 10% DIF Conditions, $n_R = n_F = 500$

Method	N(0, 1) focal group				N(-1, 1) focal group			
	Balanced: Condition 1B		Unbalanced: Condition 1D		Balanced: Condition 3B		Unbalanced: Condition 3D	
	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present
Refined	0.000	43.8	0.000	38.4	0.056	38.5	0.067	62.3
Unrefined	0.000	47.8	0.011	34.8	0.044	42.1	0.056	54.2
Average	0.000	45.8	0.006	36.6	0.050	40.3	0.062	58.3

Note. Entries in DIF absent columns are averages over 18 items. Entries in DIF present columns are averages over two items. For entries in the refined and unrefined rows, standard errors range from 0.00 to 1.58. For the average row, standard errors range from 0.00 to 1.11. DIF = differential item functioning, $n_{F\,=}$ focal group, n_R = reference group.

Table 8

DIF Flagging Rates for ETS C Rule Under 20% DIF Conditions, n_R =200, n_F = 50

Method	N(0, 1) focal group				N(-1, 1) focal group			
	Balanced: Condition 2C		Unbalanced: Condition 2E		Balanced: Condition 4C		Unbalanced: Condition 4E	
	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present
Refined	0.475	16.6	0.662	9.2	0.875	15.3	0.825	12.3
Unrefined	0.438	17.1	0.763	8.7	0.775	14.8	0.988	11.3
Average	0.457	16.9	0.713	9.0	0.825	15.1	0.907	11.8

Note. Entries in DIF absent columns are averages over 16 items. Entries in DIF present columns are averages over four items. For entries in the refined and unrefined rows, standard errors range from 0.07 to 0.84. For the average row, standard errors range from 0.05 to 0.59. DIF = differential item functioning, $n_{F\,=}$ focal group, n_R = reference group.

Table 9

DIF Flagging Rates for ETS C Rule Under 10% DIF Conditions, n_R =200, n_F = 50

Method	N(0, 1) focal group				N(-1, 1) focal group			
	Balanced: Condition 2B		Unbalanced: Condition 2D		Balanced: Condition 4B		Unbalanced: Condition 4D	
	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present
Refined	0.567	14.9	0.511	12.7	0.744	15.4	0.767	16.5
Unrefined	0.533	15.1	0.544	12.2	0.767	15.6	0.822	15.4
Average	0.550	15.0	0.528	12.5	0.756	15.5	0.795	16.0

Note. Entries in DIF absent columns are averages over 18 items. Entries in DIF present columns are averages over two items. For entries in the refined and unrefined rows, standard errors range from 0.08 to 1.17. For the average row, standard errors range from 0.05 to 0.82. DIF = differential item functioning, $n_{F\,=}$ focal group, n_R = reference group.

Collapsing over the two focal group distributions and the two levels of DIF (not shown) reveals that, on average, refined procedures yield slightly higher detection rates than unrefined procedures. Also, regardless of whether refined or unrefined methods are used, detection rates are, on average, lower in the unbalanced than in the balanced conditions.

DIF detection for positive versus negative DIF items. The balanced conditions in our study included both positive and negative DIF items. In the large-sample conditions with an N(-1,1) focal group, detection rates were very different for these two types of items, regardless of whether refined or unrefined analyses were conducted. Results for the 20% DIF conditions are shown in Table 10. In Condition 1C, where the focal group distribution was N(0,1), positive and negative DIF items had fairly similar detection rates. However, in Condition 3C, where the focal group distribution was N(-1,1) , there was a large discrepancy, particularly for the refined analysis, where the detection rate for negative DIF items was about 78%, compared to about 6% for positive DIF items. These discrepancies are directly related to the biases in *MH D-DIF* statistics described in an earlier section. As shown in Table 3, large negative biases occurred in Condition 3C, especially for the refined DIF statistics. A negative bias causes negative DIF to be inflated in magnitude, while positive DIF is reduced.

Table 10

Detection Rate for ETS C Rule for Positive and Negative DIF Items Under Balanced 20% DIF, $n_R = n_F = 500$

Method	Condition 1C: Focal distribution N(0,1)			Condition 3C: Focal distribution N(-1,1)		
	Positive DIF	Negative DIF	Average	Positive DIF	Negative DIF	Average
Refined	48.6	46.1	47.4	6.4	77.7	42.1
Unrefined	55.6	52.2	53.9	17.6	71.2	44.4
Average	52.1	49.3	50.6	12.0	74.5	43.2

Note. Each entry in the main body of the table is an average over two items. DIF = differential item functioning, $n_{F\,=}$ focal group, n_R = reference group.

Anomalous detection rates in individual items. Examination of individual item results led to some interesting discoveries. For example, we found that for certain items, the detection rate for the refined method was lower in large-sample conditions than in conditions that were identical except for *smaller* sample size. One example is an item with a true DIF value of 1.619. As shown in Table 11, the detection rate for the refined method was lower (6.0%) in Condition 3C (20% DIF in a balanced pattern, an N(-1,1) focal group, and $n_R = n_F = 500$) than in Condition 4C (8.4%), which is the same as 3C except that the sample size is $n_R = 200$, $n_F = 50$. The corresponding results for unrefined methods are also shown in Table 11, along with the average *MH D-DIF* statistic (over 500 replications) corresponding to each of the four detection rates.

Table 11

Detection Rates for ETS C Rule and Average MH D-DIF Statistics for an Item With True DIF of 1.619

Condition	Sample sizes	Refined		Unrefined	
		Detection rate	Average MH	Detection rate	Average MH
3C	$n_R = n_F = 500$	6.0	1.089	15.4	1.244
4C	$n_R = 200$, $n_F = 50$	8.4	1.239	9.2	1.291

Note. This result applies to a single item in Conditions 3C and 4C: tests with 20% balanced DIF and an N(-1,1) focal group ability distribution. DIF = differential item functioning, $n_{F\,=}$ focal group, n_R = reference group.

Although a lower detection rate in a larger sample seems impossible at first glance, the finding proved to be correct. All four *MH D-DIF* statistics were biased downward, but this was particularly true of the *MH D-DIF* statistic for the refined analysis in Condition 3C, where $n_R = n_F = 500$. Here, the *MH D-DIF* of 1.089 was far lower than the true DIF value of 1.619. An examination of the postrefinement matching criterion showed that the refined analysis tended to exclude two items with large negative DIF from the matching criterion, while an item with large positive DIF (like the studied item) tended not to be excluded. Thus, the matching criterion was systematically distorted after refinement even though the DIF on the test was balanced prior to refinement. In the condition with $n_R = 200$, $n_F = 50$, there were few deletions due to refinement, so the DIF in the matching criterion tended to be balanced. These analyses provide an illustration of why refined analyses generally work more poorly than unrefined analyses when DIF is balanced: They can disrupt the existing balance in the matching criterion.

Flagging Rates for the Chi-Square-Only Approach

As expected, both Type I error rates and DIF detection rates were much higher for the chi-square-only approach than for the ETS *C* rule. For the four no-DIF conditions, average Type I error rates are shown in Table 12. For the items without DIF in the 20% DIF conditions, Type I error rates are given in the DIF absent columns of Tables 13 and 14. The detection rates for the 20% DIF conditions are shown in the DIF present columns of these tables. Results for the 10% DIF conditions (not shown) revealed similar patterns.

Table 12

DIF Flagging Rates for Chi-Square-Only Procedure Under Conditions Without DIF

Sample size	Condition	Method	Flagging rate
$n_R = n_F = 500$	1A	Refined	3.8
		Unrefined	4.1
	3A	Refined	8.1
		Unrefined	8.0
$n_R = 200$, $n_F = 50$	2A	Refined	2.7
		Unrefined	2.9
	4A	Refined	3.4
		Unrefined	3.5

Note. DIF = differential item functioning. Each entry is an average over 20 items, with 500 replications per item. Standard errors of table entries range from 0.16 to 0.27.

Table 13

DIF Flagging Rates for Chi-Square-Only Procedure Under 20% DIF Conditions, $n_R = n_F = 500$

Method	N(0, 1) focal group				N(-1, 1) focal group			
	Balanced: Condition 1C		Unbalanced: Condition 1E		Balanced: Condition 3C		Unbalanced: Condition 3E	
	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present
Refined	3.7	97.2	4.3	92.1	9.0	89.5	8.5	98.8
Unrefined	4.0	98.5	8.7	85.2	6.9	95.5	12.6	94.3
Average	3.9	97.9	6.5	88.7	8.0	92.5	10.6	96.6

Note. Entries in DIF absent columns are averages over 16 items. Entries in DIF present columns are averages over four items. For entries in the refined and unrefined rows, standard errors range from 0.21 to 0.79. For the average row, standard errors range from 0.15 to 0.50. DIF = differential item functioning, $n_{F\,=}$ focal group, n_R = reference group.

Table 14

DIF Flagging Rates for Chi-Square-Only Procedure Under 20% DIF Conditions, $n_R = 200$, $n_F = 50$

Method	N(0, 1) focal group				N(-1, 1) focal group			
	Balanced: Condition 2C		Unbalanced: Condition 2E		Balanced: Condition 4C		Unbalanced: Condition 4E	
	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present	DIF absent	DIF present
Refined	2.9	30.8	3.4	21.4	3.6	30.4	3.7	29.4
Unrefined	2.8	33.1	3.7	19.3	3.4	31.9	4.2	25.4
Average	2.9	32.0	3.6	20.4	3.5	31.2	4.0	27.4

Note. Entries in DIF absent columns are averages over 16 items. Entries in DIF present columns are averages over 4 items. For entries in the refined and unrefined rows, standard errors range from 0.18 to 1.05. For the average row, standard errors range from 0.13 to 0.74. DIF = differential item functioning, $n_{F\,=}$ focal group, n_R = reference group.

Tables 12 through 14 show that, except in Condition 3, Type I error rates were generally conservative (i.e., less than the nominal alpha level of .05), consistent with theory and with previous findings on the MH chi-square with continuity correction (e.g., Paek, 2010). Holland and Thayer (1988, p. 135) noted that the purpose of the continuity correction is to "improve the

calculation of the observed significance levels using the chi-square table rather than to make the size of the test equal to the nominal value." (The continuity correction has no bearing on the identification of *C* items using the ETS rule, since, as noted earlier, the MH chi-square test is not used to determine this designation.)

The higher Type I error rates in Condition 3 relative to 1 and in Condition 4 relative to 2 are also consistent with theory and with previous research. As noted in the earlier discussion of bias, matching errors can occur when the two groups have different ability distributions, leading to the appearance of DIF when none exists (see Uttaro & Millsap, 1994; Zwick, 1990).

The largest Type I error rates occurred in Condition 3E (Table 13), where the groups have different ability distributions, DIF is unbalanced, and the sample size is large. Here, the Type I error rate was 12.6% for the unrefined method and 8.5% for the refined method. The refined method, despite having a lower Type I error rate, had a higher DIF detection rate. Similarly, in Condition 3C (different ability distributions, balanced DIF, large samples), the unrefined method had a considerably higher detection rate, but a lower Type I error rate than the refined method. In general, as in the case of the ETS *C* rule, unrefined methods were advantageous in balanced conditions, while refined methods were advantageous in unbalanced conditions. Refined and unrefined methods produced fairly similar flagging rates in the small-sample conditions (Table 14).

It is interesting that, as shown in Table 13, the chi-square-only approach led to almost perfect DIF detection with conservative Type I error rates in Condition 1C (same ability distribution, balanced DIF, large samples). However, Table 13 also shows that the use of the chi-square-only approach can lead to large Type I error rates, particularly when the two groups have different ability distributions.

Summary and Recommendations

The main portion of our study involved an investigation of the impact of refinement on the flagging rates obtained with the ETS *C* rule, which dictates that an item be flagged if $|MH\ D\text{-}DIF|$ is greater than 1.5 and is statistically different from 1 at $\alpha = .05$. Type I error rates for the ETS *C* rule were extremely low—less than 1% in all conditions for both refined and unrefined methods. Type I error rates tended to be higher for the refined method in balanced conditions and higher for the unrefined method in unbalanced conditions.

One of our key findings was that, for the large-*n* conditions ($n_R = n_F = 500$), the refined DIF method had a higher detection rate than the unrefined when DIF was unbalanced; the

unrefined method performed better with balanced DIF. This finding was consistent with our initial conjecture. Supplementary analyses of the ETS *C* rule (not shown) revealed that the advantage of the refined analysis in unbalanced cases was much smaller when the focal group (the group disadvantaged by DIF) had a distribution 1 standard deviation *higher* than that of the reference group. The unrefined method, however, retained its advantage in balanced cases. For example, with 20% unbalanced DIF and $n_R = n_F = 500$, the refined method had a detection rate of 10.8%, compared to 9.8% for the unrefined method. With 20% balanced DIF, the unrefined method had a detection rate of 52.6%, compared to 46.1% for the refined method.

In both the balanced and unbalanced cases, we found that the method with the higher detection rate was most often the one with the lower average bias and, unexpectedly, the lower Type I error rate. DIF detection rates for refined and unrefined methods differed only slightly in the small-*n* conditions ($n_R = 200$, $n_F = 50$). Because of low statistical power, items were unlikely to be excluded from the matching criterion in the preliminary DIF run, resulting in refined analyses that were similar to the unrefined analyses.

Some anomalous situations occurred in which refined methods produced a lower detection rate with large samples than with small samples. On average, balanced DIF conditions led to higher detection rates, regardless of whether refined or unrefined analysis was used.

The DIF flagging rates for the ETS *C* rule are, of course, partly dependent on the accuracy with which DIF is estimated. In our study, the bias in the MH statistics was substantial in some conditions. The observed patterns of bias can be attributed both to the unreliability of the matching criterion and to the self-norming nature of the MH statistic. In our simulation, the bias could be predicted almost perfectly based on the true DIF values used in data generation.

In addition to investigating the ETS *C* rule, we considered a rule in which an item was flagged if the MH chi-square statistic was statistically significant at $\alpha = .05$. The results for this rule are somewhat easier to interpret because only one criterion needs to be satisfied (i.e., effect size is not considered). Type I error rates were generally conservative, except when the reference and focal groups had different ability distributions and samples were large. As in the case of the ETS *C* rule, the refined DIF method had a higher detection rate than the unrefined when DIF was unbalanced; the unrefined method performed better with balanced DIF. In most cases, higher detection rates were, again, associated with lower Type I error rates.

Our analyses showed that in small samples, refinement does not seem advantageous. We also found that refinement had essentially no effect when the test had 80 items, even when 16 of the items had DIF. The situation was different for the 20-item tests in the large-sample case, which led to better results for the refined method in the case of unbalanced DIF and better results for the unrefined method in the case of balanced DIF. Our analyses revealed that, in both small and large samples, refinement can actually disrupt a satisfactorily balanced matching criterion (see Table 11). If previous research or theoretical considerations suggest that DIF is likely to be balanced, then the unrefined method is likely to produce better results, whereas, if unbalanced DIF is expected, the opposite is true. In the absence of information, it is probably best to choose the refined method because it is only slightly disadvantageous in balanced conditions, whereas the unrefined method can be substantially disadvantageous in certain unbalanced conditions.

Future research on refinement could seek to determine the sample sizes and test lengths at which the distinction between refined and unrefined analyses becomes important. It would also be useful to generalize the results beyond the DIF assessment methods investigated here.

References

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

Colvin, K. F., & Randall, J. (2011). *A review of recent findings on DIF analysis techniques* (Center for Educational Assessment Research Report. No. 795). Amherst: University of Massachusetts, Amherst, Center for Educational Assessment.

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

French, B., & Maller, S. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*, 373–393.

Holland, P. W., & Thayer, D. T. (1986a, April). *Differential item performance and the Mantel-Haenszel procedure.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco. Retrieved from ERIC database. (ED272577).

Holland, P. W., & Thayer, D. T. (1986b). *Differential item functioning and the Mantel-Haenszel procedure* (Research Report No. RR-86-31). Princeton, NJ: Educational Testing Service.

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

Kok, F. G., Mellenbergh, G. J., & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22*, 295–303.

Lord, F. M. (1976, July). *A study of item bias using characteristic curve theory.* Retrieved from the ERIC database. (ED137486)

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

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

Lord, F. M., & Novick, M. R. (1968*). Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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.

Paek, I. (2010). Conservativeness in rejection of the null hypothesis when using the continuity correction in the MH chi-square test in DIF applications. *Applied Psychological Measurement, 34*, 539–548.

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

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

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

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report. No. RR-12-08). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25*, 225–247.

Zwick, R., Ye, L., & Isham, S. (2012). Improving Mantel-Haenszel DIF estimation through Bayesian updating. *Journal of Educational and Behavioral Statistics, 37*, 601–629.

Notes

[1] We were able to demonstrate that the average signed bias for a simulation condition could be predicted almost perfectly using the amount of true DIF in the matching criterion. This work is available from the authors.