Listening. Learning. Leading.®

Research Report
ETS RR-12-08

A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement

Rebecca Zwick

May 2012

A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement

Rebecca Zwick
ETS, Princeton, New Jersey

May 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Marna Golub-Smith

Technical Reviewers: Neil Dorans and Tim Moses

Abstract

Differential item functioning (DIF) analysis is a key component in the evaluation of the fairness and validity of educational tests. The goal of this project was to review the status of ETS DIF analysis procedures, focusing on three aspects: (a) the nature and stringency of the statistical rules used to flag items, (b) the minimum sample size requirements that are currently in place for DIF analysis, and (c) the efficacy of criterion refinement. The main findings of the review are as follows:

- The ETS C rule often displays low DIF detection rates even when samples are large.
- With improved flagging rules in place, minimum sample size requirements could probably be relaxed. In addition, updated rules for combining data across administrations could allow DIF analyses to be performed in a broader range of situations.
- Refinement of the matching criterion improves detection rates when DIF is primarily in one direction but can depress detection rates when DIF is balanced. If nothing is known about the likely pattern of DIF, refinement is advisable.

Each of these findings is discussed in detail, focusing on the case of dichotomous items.

Key words: differential item functioning (DIF), test fairness, refinement, Mantel-Haenszel DIF, empirical Bayes DIF

Acknowledgments

I am grateful for the contributions of Kimberly Colvin of the University of Massachusetts, who conducted a literature review in support of this project (Colvin & Randall, 2011), and of Lei Ye and Steven Isham, who conducted the simulations needed to produce the tables. The DIF refinement analyses are described in full in a paper by Zwick, Ye, and Isham (2012). I also appreciate the thoughtful reviews provided by Neil Dorans, Marna Golub-Smith, Shelby Haberman, and Tim Moses.

Table of Contents

List of Tables

Page

Differential item functioning (DIF) analysis is a key component in the evaluation of the fairness and validity of educational tests. As part of its standard operations, ETS conducts DIF analyses on thousands of items per year. It is therefore important that these analyses be conducted in such a way as to produce the most accurate and useful results. The goal of this project was to investigate the status of ETS DIF analysis procedures, focusing on three aspects:

- the nature and stringency of the statistical rules used to flag items,
- the minimum sample size requirements that are currently in place for DIF analysis,
- and the efficacy of criterion refinement.

It was suggested by ETS Research management that this project could serve as a first step in a more comprehensive multiyear review of the ETS DIF policies and procedures. Although the current system has served ETS well, it is worthwhile to reexamine its provisions. The project comprised several activities: a literature review, a series of simulations, and a survey of ETS staff. This report also draws on a study conducted by Zwick, Ye, and Isham (in press) as well as other past research.

The main findings of the review are as follows:

- The ETS C rule often displays low DIF detection rates even when samples are large.
- With improved flagging rules in place, minimum sample size requirements could probably be relaxed. In addition, updated rules for combining data across administrations could allow DIF analyses to be performed in a broader range of situations.
- Refinement of the matching criterion improves detection rates when DIF is unbalanced (i.e., primarily in one direction), but can depress detection rates when DIF is balanced. If nothing is known about the likely pattern of DIF, refinement is advisable.

Each of these findings is discussed in detail below, focusing on the case of dichotomous items. The final section of the paper is a summary of recommendations and a discussion of other DIF issues that may merit further examination. The results of a 2011 survey of ETS staff are discussed in that context.

Nature and Stringency of Rules Used to Identify Differential Item Functioning Items at ETS

The ETS system for DIF classification has been in place for nearly 25 years. As described by Zieky (1993, p. 342), statistical analyses are used to designate items as A (*negligible or nonsignificant DIF*), B (*slight to moderate DIF*), or C (*moderate to large DIF*). Over the years, minor changes have been made in the statistical formulae used to assign items to these categories, but the overall classification system has remained intact. The formulae, as well as the sample size requirements for DIF analysis, are currently documented in a series of memos, dating back to a 1987 memo by Nancy Petersen, who was then a senior psychometrician at ETS.

As detailed below, the rules currently used at ETS classify items as A, B, or C items depending on the magnitude of the Mantel-Haenszel delta difference (*MH D-DIF*) statistic and its statistical significance.[1] The Mantel-Haenszel (1959) approach to DIF analysis, developed by Holland and Thayer (1988), involves the creation of *K* two-by-two tables, where *K* is the number of score categories on the matching criterion. For the *k*th score level, the data can be summarized as follows: N_{R1k} and N_{F1k} denote the numbers of examinees in the reference and focal groups, respectively, who answered correctly; N_{R0k} and N_{F0k} are the numbers of examinees in the reference and focal groups who answered incorrectly. N_k is the total number of examinees. The Mantel-Haenszel estimate of the conditional odds ratio is defined as

$$\hat{\alpha}_{MH} = \frac{\sum_k N_{R1k} N_{F0k} / N_k}{\sum_k N_{R0k} N_{F1k} / N_k}. \qquad (1)$$

The corresponding population parameter, α_{MH}, is assumed to be constant over all levels of the matching criterion.

The *MH D-DIF* index, which was developed by Holland and Thayer (1988), is defined as follows:

$$MH\ D\text{-}DIF = -2.35 \ln(\hat{\alpha}_{MH}). \qquad (2)$$

By convention, $\hat{\alpha}_{MH}$ is formulated so that MH-D-DIF is negative when the focal group odds of correct response are less than the reference group odds, conditional on the matching variable.

In developing the *MH D-DIF* index, Holland and Thayer (1985) elected to express the statistic on the ETS delta scale of item difficulty. An *MH D-DIF* value of -1, for example, means that the item is estimated to be more difficult for the focal group than for the reference group by an average of one delta point, conditional on ability. Expressing the amount of DIF in this way was intended to make the *MH D-DIF* statistic more useful for ETS test developers.

For those who prefer to think in terms of odds ratios, an *MH D-DIF* statistic of -1 implies that -2.35 $\ln \hat{\alpha}_{MH}$ = -1, or $\hat{\alpha}_{MH}$ = 1.530. This means that the odds of answering correctly for the reference group are more than 50% higher than the odds of answering correctly for comparable members of the focal group. (An *MH D-DIF* of +1 means that the odds of answering correctly for the reference group are 1/1.530 = .653 times the odds of answering correctly for comparable members of the focal group.)

Detailed Analysis of A, B, and C Categories

As noted earlier, ETS classifies DIF items into three categories: A, B, and C. Items labeled B and C are further distinguished by their signs: B+ and C+ items are those that show DIF in favor of the focal group; B- and C- items show DIF in favor of the reference group. For the purpose at hand, it is useful to first define an A item, then a C, and last, a B.

An A item is one in which either the Mantel-Haenszel (MH) chi-square statistic is not significant at the 5% level or *MH D-DIF* is smaller than 1 in absolute value. The MH chi-square statistic, as implemented at ETS, is defined as follows:

$$MH\ CHISQ = \frac{(|\sum_k N_{R1k} - \sum_k E(N_{R1k})| - \frac{1}{2})^2}{\sum_k Var(N_{R1k})}, \qquad (3)$$

where $E(N_{R1k}) = n_{Rk} m_{1k} / N_k$, $Var(N_{R1k}) = \frac{n_{Rk} n_{Fk} m_{1k} m_{0k}}{N_k^2 (N_k - 1)}$, n_{Rk} and n_{Fk} denote the numbers of examinees in the reference and focal groups, respectively, m_{1k} represents the number of examinees who answered the item correctly, and m_{0k} is the number who answered incorrectly. The statistic in Equation 3 has a chi-square distribution with one degree of freedom when the null hypothesis of a constant odds ratio equal to one is true. (The ½ that is subtracted in the

numerator is a continuity correction, designed to improve the approximation of a discrete distribution with a continuous distribution. It is discussed further in a later section.)

In order to qualify as a C item, the *MH D-DIF* statistic must be significantly greater than 1 in absolute value at the 5% level and must have an absolute value of 1.5 or more. The hypothesis testing procedure is complicated by the fact that there is a composite null hypothesis—that is, the null hypothesis corresponds to a region (between -1 and +1), not a point. In an internal ETS memo, Paul Holland (2004) showed that the correct critical value is the 95th percentile of the standard normal distribution, which is 1.645. Therefore, an item is classified as a C if

$$(|MH\ D\text{-}DIF|-1)/SE(MH\ D\text{-}DIF) > 1.645, \quad (4)$$

where $SE(MH\ D\text{-}DIF)$ is the estimated standard error of *MH D-DIF*, and if

$$|MH\ D\text{-}DIF| \geq 1.5. \quad (5)$$

It is worth noting that, if $SE(MH\ D\text{-}DIF) \leq .304$, then the statistical significance criterion in Equation 4 is superfluous because any item that meets the criterion in Equation 5 will also meet the criterion in Equation 4.

Items that do not meet the definition for either A or C items are considered B items. More explicitly, an item is declared a B item if it does not meet the qualifications for a C item and if the following two conditions hold:

$$MH\ CHISQ > 3.84 \quad (6)$$

and

$$|MH\ D\text{-}DIF| \geq 1. \quad (7)$$

Holland (2004) noted that the rule in (6) is asymptotically equivalent to the rule,

$$(|MH\ D\text{-}DIF|)/SE(MH\ D\text{-}DIF) > 1.96. \quad (8)$$

We can show that if $SE(MH\ D\text{-}DIF) < .510$, then the statistical significance criterion in Equation 8, which is roughly equivalent to the rule in Equation 6, will be satisfied by any *MH-D-DIF*

value that satisfies Equation 7. That is, if *SE*(*MH D-DIF*) is small enough, then any *MH-D-DIF* with an absolute value of 1 or more will satisfy Equation 8. [2]

According to Dorans (personal communication, August 16, 2011), the reasoning behind the cutoffs of 1 and 1.5 embedded in these definitions was as follows: "A delta difference of 1 point, while undesirable, can be tolerated. . . . A difference of 2 or more, however, should be avoided. An unrounded delta difference of 1.5 represents the lower limit of a delta of 2.0 (1.5 to 2.5)."

In the next section of this review, some simulation findings on the effectiveness of alternative flagging rules are presented, followed by some considerations for modifying the current ETS rules.

Effectiveness of Three Rules for Flagging Differential Item Functioning Items

Let us examine the effectiveness of several rules for identifying items with DIF, based on the results of simulation studies. Table 1 shows results for each item with DIF exceeding 1 in the *MH D-DIF* metric[3] that were included in a simulation conducted by Zwick et al. (in press). The table includes conditions in which the reference and focal groups each had 500 members, well above ETS sample size criteria. Column 1 gives the magnitude of DIF in the *MH D-DIF* metric. Columns 2 and 3 give the DIF detection rates for the ETS C rule. The column labeled *same* pertains to results obtained when the reference and focal group have the same ability distribution (standard normal); the column labeled *different* pertains to results obtained when the focal group distribution is one standard deviation lower than the reference group distribution. In general, DIF methods that use observed-score matching perform more poorly when the groups have different ability distributions because matching tends to be less accurate (see Uttaro & Millsap, 1994; Zwick, 1990). The next pair of columns provides results for the rule that flags items that are at least a B (i.e., items that are B's or C's). The third pair of columns provides results of a flagging approach originally presented in Zwick, Thayer, and Lewis (2000) and applied in modified form by Zwick et al. (in press). This rule, along with the results in the last two columns, labeled *revised ETS rule*, is discussed further below. For each rule, the average detection rate for the 13 items with DIF exceeding 1 in magnitude is given.

The last row of the table shows, for each DIF rule, the average rate of DIF identification for the 21 items in the simulation that had only negligible DIF (less than 1 in magnitude). If the

Table 1

Flagging Rates for Differential Item Functioning Items in Simulated Data With 500 Members per Group

Simulated DIF in *MH D-DIF* metric	DIF flagging rule							
	C rule		B rule (Flag if B or C)		EB loss function conservative rule		Revised ETS rule	
	Group ability distributions							
	Same	Different	Same	Different	Same	Different	Same	Different
-2.64	96.6	99.8	100.0	100.0	99.8	100.0	100.0	100.0
-1.85	82.0	59.2	99.8	97.0	95.4	82.8	97.2	90.4
-1.36	35.6	15.2	88.6	72.6	69.8	37.8	77.2	52.6
-1.23	12.6	31.0	68.2	87.4	37.8	60.8	49.6	73.2
-1.10	10.2	4.6	69.4	45.8	33.2	18.8	42.8	25.8
-1.04	7.4	3.0	58.0	42.4	25.2	12.6	31.8	21.2
1.02	13.2	5.2	70.6	45.6	40.2	16.2	49.4	30.8
1.18	12.0	2.4	64.6	37.4	36.0	13.4	42.6	20.2
1.23	19.2	12.4	80.2	70.0	50.4	37.6	59.6	50.2
1.80	82.4	37.2	99.6	88.0	95.0	68.2	97.0	77.8
1.83	57.4	19.8	96.4	81.0	85.6	47.6	90.6	62.4
2.03	43.2	11.4	86.2	63.0	79.4	36.6	85.6	51.4
3.12	100.0	97.2	100.0	100.0	100.0	100.0	100.0	100.0
Average flagging rate for above items (true B and C items)	44.0	30.6	83.2	71.6	65.2	48.6	71.0	58.2
Average flagging rate for 21 items (not shown) with simulated DIF < 1 (true A items)	0.4	0.3	10.8	8.4	3.1	2.2	4.8	4.6

Note. No refinement was used in these analyses. Individual item results are based on 500 replications. The average flagging rate for true A items is based on 21 items, with 500 replications per item. The C rule is identical to the ETS C rule. The B rule flags items that are B or C according to the ETS criteria. The EB loss function conservative rule is a variation on a rule developed by Zwick et al. (2000). Details on the rules and the simulation procedures are given in Zwick et al. (in press). DIF = differential item functioning, EB = empirical Bayes, *MH D-DIF* = Mantel-Haenszel delta difference.

null hypothesis is assumed to correspond to the region between -1 and +1 in the *MH D-DIF* metric, as described by Holland (2004), this is an average Type I error rate and will be referred to as such here.[4] One would like this rate to be low. On the other hand, if one is willing to tolerate only a very low level of false identification, the power to detect existing DIF will also be low. This is the usual trade-off between Type I error (rejecting the null hypothesis when it is true) and Type II error (failing to reject the null hypothesis when it is false). In the DIF context, Type I error may be of less concern than in more conventional hypothesis testing situations: From the perspective of equity in assessment, the costs of falsely identifying an item as having DIF are low, while the costs of failing to identify a DIF item are high. (This is the rationale that has been given for not imposing a Bonferroni-type correction to control the overall Type I error rate in ETS DIF assessment.)

The rules in Table 1 differ substantially in terms of their Type I error rates. Consider the C and B rules. Under the C rule, it is very rare for an item with negligible DIF (an A item) to be mistakenly flagged: These false positives occur less than one half of one percent of the time. Under the B rule (which flags B and C items), the average flagging rate for A items is 8% to 11%. The DIF identification rates parallel the Type I error rates: For the C rule, the average identification rate is 44% when ability distributions for the two groups are the same and 31% when they are different. The detection rate for most items is less than 50%. This includes some items with substantial DIF. For example, an item with DIF of 2.03 in the MH metric is flagged only 11% of the time when the reference and focal group distributions differ by one standard deviation. (The result for this item is a good illustration of the fact that flagging rates are not a simple function of the true DIF values; item difficulty and discrimination play a role as well.) For DIF values close to 1, the identification level falls as low as 2.4%. The B rule, by contrast, always has flagging rates of at least 58% when the two groups have the same ability distribution, with an average rate of 83%. When the groups have different ability distributions, the rates fall below 50% for four items, reaching a minimum of 37% for an item with DIF of 1.18. The average rate in this condition is 72%.

The rule labeled *EB loss function conservative rule* in Table 1 does not use either statistical significance testing or minimum values for the magnitude of *MH D-DIF*. The rule is based on a Bayesian approach in which the distribution of a presumed DIF parameter is estimated. The decision about whether to flag an item is based on a loss function. The empirical

Bayes (EB) procedures were developed by Zwick, Thayer, and Lewis (1999, 2000) based on suggestions by Holland (1987a, 1987b) and others. In the EB approach, a prior distribution for the DIF parameter ω is assumed. MH statistics and their standard errors from the current test administration are used as a basis for estimating the mean and variance of the prior. Because the prior distribution and the likelihood function are both assumed normal, the posterior distribution of ω is also normal and its mean serves as the EB estimate of DIF. The EB approach was found to produce more stable DIF estimates than the ordinary MH method. Also, a loss-function-based DIF detection rule that made use of the EB results was often better able to identify DIF items than the ABC classification system. The particular rule in Table 1 is a modified version of the original rule developed by Zwick et al. (2000; see also Sinharay, Dorans, Grant, & Blew, 2009). The results of this modified rule (detailed in Zwick et al., in press) illustrate that approaches other than significance testing and effect size criteria can yield good results. Although the EB rule has Type I error rates higher than those of the C rule, its error rates of 2% to 3% are well below conventional levels. The EB rule has flagging rates much higher than the C rule. Consider the item with DIF of 2.03 in the condition where the reference and focal groups have the same distribution. The EB rule flags this item 79% of the time, compared to 43% for the C rule.

Table 2 shows detection rates and Type I error rates for conditions in which the reference group had 200 members and the focal group had 50 members. These sample sizes do not meet ETS guidelines for DIF analysis. As expected, detection rates are much lower than for the large-sample conditions of Table 1. Again, the C rule has very low Type I error rates but has detection rates averaging only 18% and 13% in the same and different conditions, respectively. The B rule does much better, with Type I error rates of around 5% and detection rates of 32% and 25% for the same and different conditions. Although the EB estimates themselves performed well in the small-sample conditions (i.e., they had substantially smaller average departures from their target values than the *MH D-DIF* statistics), the loss-function based rule was insufficiently stringent here, at least in the same condition, where the average Type I error rate was 16%. The average Type I error rate was 10% for the different condition. Detection rates averaged 48% and 30% in the same and different conditions, respectively.

Table 2

Flagging Rates for Differential Item Functioning Items in Simulated Data With 200 Members in the Reference Group and 50 Members in the Focal Group

Simulated DIF in *MH D-DIF* metric	DIF flagging rule					
	C rule		B rule (Flag if B or C)		EB loss function-conservative rule	
	Group ability distributions					
	Same	Different	Same	Different	Same	Different
-2.64	40.4	53.2	51.8	71.2	70.0	66.2
-1.85	28.6	16.0	54.6	33.6	62.0	40.8
-1.36	11.2	8.2	31.0	19.0	44.8	24.4
-1.23	7.0	9.8	12.0	22.6	32.2	31.4
-1.10	7.6	3.4	21.4	11.6	34.4	17.8
-1.04	6.6	3.8	18.8	12.4	30.4	18.2
1.02	7.2	4.2	18.2	10.4	33.4	17.2
1.18	4.6	3.0	13.4	7.6	30.6	14.8
1.23	7.4	6.6	24.4	16.4	38.8	21.8
1.80	27.2	10.0	52.2	25.0	63.4	29.4
1.83	16.6	8.8	29.2	21.4	51.4	27.4
2.03	3.8	5.4	5.2	10.4	42.4	21.8
3.12	61.2	35.0	78.8	57.0	86.4	54.8
Average flagging rate for above items (true B and C items)	17.6	12.9	31.6	24.5	47.7	29.7
Average flagging rate for 21 items (not shown) with simulated DIF < 1 (true A items)	1.6	1.4	5.2	4.5	15.9	10.0

Note. No refinement was used in these analyses. Individual item results are based on 500 replications. The average flagging rate for true A Items is based on 21 items, with 500 replications per item. The C rule is identical to the ETS C rule. The B rule flags items that are B or C according to the ETS criteria. The EB loss function conservative rule is a variation on a rule developed by Zwick et al. (2000). Details on the rules and the simulation procedures are given in Zwick et al. (in press). DIF = differential item functioning, EB = empirical Bayes, *MH D-DIF* = Mantel-Haenszel delta difference.

Recommendations—Developing More Effective Differential Item Functioning Flagging Rules

To evaluate the adequacy of a decision rule, it is necessary to have a goal in mind. For example, *the DIF rule* should correctly detect DIF exceeding 1.5 in magnitude at least 50% of the time, with a Type I error rate averaging no more than 5% per item. Table 1 shows that with 500 members per group, even this modest goal is not close to being attained by the C rule; it is closer to being met by the B rule and the EB rule. Clearly, the C rule optimizes Type I error control to the detriment of detection. (Tables 5 and 6, which are discussed in the section on refinement, provide further evidence of the C rule's low detection rate.)

Two possible directions for devising more effective DIF rules are (a) the development of a rule, similar in form to the current ETS rules, that is less stringent than the C rule but somewhat more stringent than the B rule and (b) the development of a loss function-based rule that results in smaller Type I error rates for conditions similar to those in Table 2. For either type of rule, the first step should be to reconsider the issue of minimal DIF magnitude that is of concern (and is therefore important to detect) as well as the level of false positives that can be tolerated. (As discussed further in the overall summary and discussion, this minimal DIF magnitude need not be expressed in the delta metric.) Then, a combination of theoretical findings and simulations can be used to develop a rule that is consistent with the goals that have been defined.

As one example of a possible alternative flagging procedure, consider a rule that flags items if the *MH CHISQ* statistic is significant at the 5% level and the absolute value of the *MH D-DIF* statistic is at least 1.2. The results of this rule for DIF analyses with 500 members per group are given in the two right-most columns of Table 1, under the heading labeled *revised ETS rule*. For items with true DIF of 1.5 or more, the detection rates range from 51.4% to 100%. The average Type I error rates are slightly under 5%. When the reference group had 200 members and the focal group had 50 members, results were identical to those for the B rule in Table 2. (*MH D-DIF* values between 1 and 1.2 always led to nonsignificant chi-square values and therefore did not lead to DIF flagging under either the B rule or the revised ETS rule.)

One question that merits further attention is whether it is useful to maintain three categories of DIF severity (A, B, and C). ETS guidelines for test development incorporate all three categories. A 1988 memorandum states that, in general, "Items from Category A should be selected in preference to items from Categories B or C. . . . For items in Category B, *when there*

is a choice among otherwise equally appropriate items, then items with smaller absolute *MH D-DIF* values should be selected. . . . Items from Category C will NOT be used unless they are judged to be essential to meet test specifications" (Educational Testing Service, 1988, p. 8; emphasis in original). Although this guideline still plays a role in assessment development, it is the C rule that is typically used for purposes of identifying DIF items for review by committees or possible deletion. Only limited attention is paid to the B category in this context. Given the ambiguous status of the B category, it may be advisable to consider a binary classification system.

Sample Size Requirements for ETS Differential Item Functioning Analysis

The sample size requirements for ETS DIF analysis are currently documented in a series of memos, dating back to Petersen's 1987 memo. Currently, ETS programs that do not meet sample size requirements for certain pairs of groups are exempt from the requirement to perform DIF analysis for those pairs of groups. According to a 2001 memo from senior ETS research directors Tim Davey and Cathy Wendler, at least 200 members in the smaller group and at least 500 in total are needed for DIF analyses performed at the test assembly phase. For DIF analyses performed at the preliminary item analysis phase (after a test has been administered but before scores are reported), the minimum sample size requirements are 300 members in the smaller group and 700 in total. The rationale for the sample size requirements is that analysis results are likely to be unstable with smaller samples.

Among the approaches that have been proposed in the research literature to enhance the utility of MH DIF detection methods in small samples are (a) exact, jackknife, bootstrap, or randomization-based versions of the MH method; (b) Bayesian modifications of the MH procedures; (c) the use of large nominal Type I error rates with the MH chi-square test; (d) elimination of the continuity correction in the MH chi-square test; and (e) aggregation of DIF information across multiple administrations (or administration windows). Each of these proposals is discussed below.

The *MH CHISQ* statistic is approximately distributed as chi-square with one degree of freedom under the null hypothesis. Because of a concern that the approximation may be inadequate in small samples, some DIF researchers have proposed analogues to the MH chi-square test that do not rely on large-sample approximations. Camilli and Smith (1990) applied a randomization-based approximation to the exact permutation test corresponding to the MH test, as well as a procedure in

which the log of the MH odds ratio estimate is divided by a jackknife estimate of its standard error. Their analyses were based on real and simulated data with a reference group sample size of 1,085 and a focal group sample size of 300. They found that the alternative statistical approaches led to essentially the same results as "the unadorned [continuity-corrected] MH chi square" (Camilli & Smith, p. 63). Parshall and Miller (1995) compared the MH chi-square (without continuity correction) to a procedure based on the exact permutation test. The reference group sample size was 500 and the focal group sample size ranged from 25 to 200. The authors concluded that the "exact methods offered no particular advantage over the asymptotic approach under small-sample conditions" (p. 311). Similarly, Lu and Dunbar (1996) found that a bootstrap version of the MH yielded results that were very similar to those of the standard procedure even when the focal group sample size was less than 100.

Bayesian elaborations of MH DIF analysis were developed by Zwick et al. (1999, 2000; see also Zwick & Thayer, 2002); modified versions of these procedures were studied by Sinharay et al. (2009) and Zwick et al. (in press).[5] These researchers found that, in general, Bayesian DIF statistics were more stable than the *MH D-DIF* statistic in small samples. For example, Zwick et al. (1999) examined root mean square residuals (*RMSRs*) that compared EB DIF statistics and *MH D-DIF* statistics to their true values in a simulation study. They found that for "samples of 200 reference group members and 50 focal group members, the behavior of the EB point estimates was substantially superior to that of *MH D-DIF*. On the average, the values of *MH D-DIF* differed from the true DIF values by about 1 in the MH metric; the median *RMSR* for the EB estimates was .65" (p. 18).

In an investigation that included the EB procedures developed by Zwick et al. (1999, 2000), Fidalgo, Hashimoto, Bartram, and Muñiz (2007) found that on average, the EB estimates had smaller *RMSRs* than *MH D-DIF* in each of 10 simulation conditions. The difference in average *RMSRs* was largest when the reference and focal groups both had 50 members. In this case, the average *RMSR* for the EB estimates was .85 in the MH metric, compared to 1.35 for *MH D-DIF* (p. 310). Nevertheless, Fidalgo et al. came to a negative conclusion regarding the EB procedures, stating that the greater stability of the EB estimator was "limited by its considerable bias" (p. 309). Fidalgo et al. also critiqued the loss-function-based approach used by Zwick et al. (2000), apparently not recognizing the fact that, like any flagging procedure, it can be made more or less stringent (e.g., see Zwick et al., in press).

Fidalgo and his colleagues have advocated using a nominal Type I error rate of .20 when performing a MH chi-square test in small samples. It is, of course, more likely that DIF will be detected with $\alpha = .20$ than with a more conventional alpha level, but high Type I error rates will be an inevitable result as well. In their simulation study, Fidalgo, Ferreres, and Muñiz (2004, p. 932) found that the empirical Type I error rate ranged from .13 to .17 for this procedure for combined sample sizes ranging from 100 to 250. When both groups had 500 members, Fidalgo et al. (2007, p. 308) found Type I error rates as high as .27.

Paek (2010), noting the conservative nature of the MH chi-square with continuity correction (see Equation 3), suggested that the continuity correction be abandoned. His simulation, which confirms earlier investigations, shows that the rejection rate for the continuity-corrected chi-square is less than the nominal value in the null case. The chi-square without continuity correction produces Type I error rates closer to the nominal value and has a lower Type II error rate. The difference between the two versions of the MH chi-square is particularly notable when samples are small. For example, when both groups had 100 members, the corrected chi-square, performed at $\alpha = .05$, had an error rate of approximately .03, while the uncorrected chi-square had an error rate essentially equal to the nominal level.

However, in arguing for the use of the continuity correction in the MH chi-square, Holland and Thayer (1988) noted that the "effect 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. Hence simulation studies routinely find that the actual size of a test based on [the corrected version] is *smaller* than the nominal value. . . . The continuity correction is simply to improve the approximation of a discrete distribution . . . by a continuous distribution" (p. 135).[6] Also, as regards ETS DIF procedures, it is important to recognize that the continuity correction has no bearing on the identification of C items, because this identification is based on Equations 4 and 5, rather than on *MH CHISQ*. The decision about whether to incorporate the continuity correction *does* affect the determination of whether an item is an A or a B item (see Equation 6).

Another way to address the problem of small sample sizes in DIF analysis is to combine data from multiple administrations. A 2001 memo from research directors Tim Davey and Cathy Wendler provided the following advice: "If necessary, pool data from two consecutive administrations within the same 12-month period in order to meet the minimum sample size requirements." This approach

(pooling the data and then performing the usual analyses) was formally investigated by Zwick et al. (in press), who labeled it the *combined-data MH* method. They compared it to another approach, which has apparently never been applied in practice, the *average MH method*. For each item, this approach uses the weighted average of *MH D-DIF* statistics from multiple administrations, as well as the standard error of that average, to classify the item into the three ETS categories. As a third approach to combining MH results, Zwick et al. (in press) introduced the Bayesian updating (BU) method, which is a multiple-administration version of the EB method described above. In the BU method, the item's DIF history, as well as its current MH results, is used in determining whether the item should be flagged. The flagging rules are based on loss functions, as in the EB method. All three methods of combining MH results appear to hold promise. The combined-data MH and average MH approaches performed very similarly. The BU approach usually performed similarly to or somewhat better than the other two approaches (Zwick et al., in press, Table 7).

Recommendations—Sample Size Requirements

Two useful directions for improving DIF detection in small samples are (a) the investigation of whether revised flagging rules could yield acceptable results in samples smaller than the current ETS minimums and (b) the reconsideration of rules and analysis procedures for combining data across multiple administrations (or administration windows). Each of these options is discussed below.

Further research could be conducted to explore the range of sample sizes for which certain DIF rules are likely to be effective. Theoretical research on this topic is possible, but given the many variables that could be manipulated (group distributions, number of test items, prevalence, size of DIF, etc.), some simulation research is inevitable. It seems likely that adjusting the flagging rules would allow the current minimum sample size rules to be relaxed. As one example, consider Table 3, which gives the results of applying the revised ETS rule in Table 1 to a data set in which $n_R = 500$, $n_F = 100$. Although the sample sizes do not meet ETS guidelines, the detection rates (an average of 59% for the same-distribution case and 46% for the different-distribution case) are far higher than those obtained using the ETS C rule when $n_R = n_F = 500$. The Type I error rates are also much higher (averaging roughly 9% and 7%, respectively for the same- and different-distribution cases), but might be considered acceptable. (Again, note that the definition of Type I error used here differs from the conventional definition. Conventional Type I error rates would be lower.)

Table 3

Flagging Rates for Differential Item Functioning Items in Simulated Data With 500 Members in the Reference Group and 100 Members in the Focal Group

Simulated DIF in *MH D-DIF* metric	Revised ETS rule	
	Ability distributions	
	Same	Different
-2.64	90.6	97.0
-1.85	88.8	70.2
-1.36	59.2	39.0
-1.23	33.8	44.8
-1.10	48.4	29.8
-1.04	39.2	27.4
1.02	41.0	20.6
1.18	36.0	18.2
1.23	55.0	39.4
1.80	83.4	52.0
1.83	60.6	44.8
2.03	31.0	24.2
3.12	98.8	91.2
Average flagging rate for above items (true B and C items)	58.9	46.0
Average flagging rate for 21 items (not shown) with simulated DIF < 1 (true A items)	9.3	6.7

Note. No refinement was used in these analyses. Individual item results are based on 500 replications. The average flagging rate for true A items is based on 21 items, with 500 replications per item. DIF = differential item functioning, EB = empirical Bayes, *MH D-DIF* = Mantel-Haenszel delta difference.

Further research on the aggregation approaches would also be useful, as well as some reconsideration of the current guidelines about combining data for DIF analyses. For example, allowing data to be pooled over a 24-month interval, rather than a 12-month interval, could be considered. Also, provisional DIF results based on small samples could be obtained and then aggregated later to obtain more stable results.

Criterion Refinement in Differential Item Functioning Analyses

Many ETS testing programs make use of criterion refinement procedures in conducting DIF analyses. Refinement is intended to improve the quality of the matching variable by removing items identified as having DIF in a preliminary round of analysis. As implemented at ETS (in GENASYS and, for NAEP, in the NDIF program), refinement involves the performance of two rounds of DIF analysis. In the second round, items that were classified as C items in the first round are deleted from the matching criterion. (An exception to this is that the studied item itself is always included.)

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. The recommendation to use criterion refinement when applying the MH DIF procedure was made by Holland and Thayer (1986a, 1986b, 1988, p. 42). Holland and Thayer 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 ETS DIF policy memos as early as 1987 (Petersen, 1987) and was repeated by Dorans and Holland (1993, pp. 60–61).

Some recent findings, however, did not support the use of refinement. In the course of a larger study, Zwick et al. (in press) 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 a simulation study that led them to conclude 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.

The Zwick et al. (in press) refinement analyses were in some ways similar to those of Clauser et al. (1993). As in their study, we simulated three-parameter logistic (3PL) data and modeled DIF as a difference between reference and focal group difficulty parameters. In addition, our surprising refinement results were based on a simulation condition similar to one

included in Clauser et al. There are many possible reasons for the discrepancy in conclusions. For example, Clauser et al. were not investigating the ETS DIF criteria but were looking only at whether the MH chi-square statistic was statistically significant at the .01 level. This was also the criterion for deleting items from the matching variable at the second stage, so the refinement process itself differed from the ETS procedure. One difference between the analyses that seems especially relevant is the pattern of the DIF that was modeled. In the Zwick et al. simulation, the differences in reference and focal group difficulties had an average near zero across the 34 items: 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. If DIF is balanced, the "contaminated" matching variable that is used in an unrefined analysis may 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 variable, degrading the results. In the Clauser et al. study, however, the unrefined matching variable had a systematic bias against the focal group members that was reduced by refinement. The disparity in results between the Clauser et al. study and our own 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 DIF flagging rules that involve both effect size and statistical significance—the rules used at ETS. Also, in evaluating the simulation outcomes, we examined the properties of the unrefined and refined MH estimates (variance, bias, root mean square residual) in addition to the Type I rate and power associated with the unrefined and refined flagging procedures. A brief description of our analyses and key results appear here. Further detail appears in Zwick, Ye, and Isham (2012).

Method

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

- Length of test (20 or 80 items)
- Percentage of items on the test with DIF (0%, 10%, or 20%); the remaining items had true DIF values of 0

- 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—against the focal group)
- Reference and focal group sample sizes ($n_R = n_F = 500$ or $n_R = 200$, $n_F = 50$)
- 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).

Item responses were generated using the 3PL model, with 500 replications per item per condition. As a starting point, we used a subset of the items (i.e., the triples of item parameters) used by Clauser et al. (1993). To induce DIF, we added or subtracted .6 from the focal group difficulty parameter. The true DIF, expressed in the MH metric (see Note 3), ranged from 0 to 2.4 in magnitude across the conditions in our study. Items with true DIF of at least 1.5 in magnitude were considered true C items.

DIF analyses were conducted with and without refinement. Our refinement procedure was identical to that used operationally at ETS: An initial DIF run was conducted, after which items identified as C items were deleted from the matching criterion. (An exception to this is that the studied item itself is always included in the matching criterion.) A second DIF run was conducted to obtain the final results.

Results

Our key findings are listed below. Further details follow.

1. Type I error rates were extremely low and were generally similar for refined and unrefined MH methods.
2. DIF detection rates for refined and unrefined methods were generally similar in the small-*n* condition *(*$n_R = 200$, $n_F = 50$). Because of low statistical power, items were unlikely to be excluded from the matching variable in the preliminary DIF run, resulting in refined analyses that were similar to the unrefined analyses.
3. Differences in detection rates between refined and unrefined methods were small in the 80-item tests, even when 20% of the items had DIF, apparently because the

number of non-DIF items (always at least 64) was sufficient to allow for reasonably accurate matching.

4. For the large-*n* conditions ($n_R = n_F = 500$) and a test length of 20, the refined DIF method had a higher detection rate than the unrefined with unbalanced DIF; the unrefined method performed better with balanced DIF. This finding was consistent with our initial conjecture.
5. In the 20-item tests, some anomalous situations occurred in which refined methods produced a lower detection rate with large samples than with small samples.

Tables 4 and 5 provide some results for the conditions that revealed the greatest differences between the refined and unrefined methods: 20-item tests with four DIF items each and large sample size ($n_R = n_F = 500$). There were four conditions with these characteristics. They varied in terms of DIF pattern and focal group distribution. In the tables, results for balanced conditions are compared to results for unbalanced conditions. Results are combined across the two focal group distributions. The true MH values for the four DIF items in the balanced conditions were 1.62, -1.63, 1.75, and -1.75. For the unbalanced conditions, they were -1.58, -1.63, -1.72, and -1.75. The amount of DIF for the balanced and unbalanced conditions is roughly equivalent in terms of the absolute magnitude of the true MH values.

Table 4

Average Squared Bias, Variance, and Root Mean Square Residuals of Mantel-Haenszel Statistics Under Balanced and Unbalanced Patterns of Differential Item Functioning

Statistic	Balanced		Unbalanced	
	Refined	Unrefined	Refined	Unrefined
Average squared bias	.0510	.0281	.0539	.0900
Average variance	.1412	.1398	.1413	.1397
Average *RMSR*	.4224	.4018	.4369	.4738

Note. Tests had 20 items with four DIF items, $n_R = n_F = 500$. Each entry is an average over 20 items, with a total of 1,000 replications per item. DIF = differential item functioning, RMSR = root mean square residuals.

Table 5

Differential Item Functioning Detection Rate With ETS C Rule Underbalanced and Unbalanced Patterns of Differential Item Functioning

DIF procedure	Balanced	Unbalanced	Average
Refined	44.7 (0.8)	33.5 (0.7)	39.1 (0.5)
Unrefined	49.2 (0.8)	25.4 (0.7)	37.3 (0.5)
Average	46.9 (0.6)	29.5 (0.5)	

Note. Tests had 20 items with four DIF items, $n_R = n_F = 500$. Each entry in the balanced and unbalanced columns is an average over 4 items and each entry in the average column is an average over 8 items, with a total of 1,000 replications per item. Average Type I error rates were near zero for both DIF procedures. Standard errors of percentages are shown in parentheses. DIF = differential item functioning.

Table 4 shows the average squared bias ($B^2(\hat{\omega})$), variance ($Var(\hat{\omega})$), and root mean square residual ($RMSR(\hat{\omega})$) of the MH statistics under balanced and unbalanced patterns of DIF. These quantities are defined as follows for each item (with item subscripts omitted for simplicity):

$$B^2(\hat{\omega}) = (\overline{\hat{\omega}} - \omega)^2, \tag{9}$$

$$Var(\hat{\omega}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\omega}_r - \overline{\hat{\omega}})^2}, \tag{10}$$

and

$$RMSR(\hat{\omega}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\omega}_r - \omega)^2} = \sqrt{B^2(\hat{\omega}) + Var(\hat{\omega})}, \tag{11}$$

where $\hat{\omega}_r$ represents the *MH D-DIF* statistic for replication *r*, $\overline{\hat{\omega}} = \frac{1}{R}\sum_{i=1}^{R}\hat{\omega}_r$ is the average of $\hat{\omega}_r$ across replications, ω is the true DIF value, and *R* is the number of replications. To construct Table 4, averages were then taken across the 20 items in a test and across the conditions corresponding to the two focal group distributions.

The average variances $Var(\hat{\omega})$ of the *MH D-DIF* statistics do not differ much across the conditions and DIF methods (refined versus unrefined), but the average squared biases $B^2(\hat{\omega})$ do. The refined method has an average $B^2(\hat{\omega})$ of about .05 in both the balanced and unbalanced conditions. The unrefined method, however, is somewhat less biased than the refined method in the balanced conditions (average $B^2(\hat{\omega})$ of .03) and more biased than the refined method in the unbalanced conditions (average $B^2(\hat{\omega})$ of .09). The $RMSR(\hat{\omega})$ values follow a similar pattern.

Table 5 shows the average DIF detection rate for each method and condition. (In these conditions, both the refined and unrefined methods had Type I error rates of zero or near-zero for the 16 non-DIF items that were included in each simulated test.) All the items included in the table were true C items. The detection rate for these items was defined as the number of replications in which the item was identified as a C using the ETS criteria, divided by the total number of replications (1,000 per item, since two conditions are combined in Table 5). The table shows that in the balanced conditions, the unrefined method has a somewhat higher detection rate (49.2% versus 44.7%), while in the unbalanced conditions, the refined method has a higher detection rate (33.5% versus 25.4%). In both the balanced and unbalanced cases, the method with the higher detection rate is the one with the lower average bias.

Another notable finding is that regardless of whether refined or unrefined DIF methods are used, detection rates are much lower in the unbalanced than in the balanced conditions (29.5% versus 46.9%). This finding is probably the result of inadequate matching in the unbalanced conditions, even after refinement is applied.[7]

Examination of individual item results led to some interesting discoveries. For example, we found that for a particular item with a true DIF value of 1.62, the detection rate for the refined method was lower in one of the large-sample conditions (6.0%) than in a condition that was identical except for smaller sample size (8.4%). These detection rates, as well as the corresponding results for unrefined methods, are shown in Table 6. Although a lower detection rate in a larger sample seems impossible at first glance, the finding proved to be correct. Whereas unrefined analysis led to correct identification in 77 of 500 replications (a detection rate of 15.4%) for $n_R = n_F = 500$, refinement produced an unbalanced matching variable—and a correspondingly lower detection rate—in 70 of these replications. Specifically, 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 variable was systematically distorted after refinement even though the DIF on the test was balanced prior to refinement. (There were no replications in which the refined analysis led to a correct identification, but unrefined analysis did not.) 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 refinement methods generally work more poorly than unrefined methods when DIF is balanced: They can disrupt the existing balance in the matching criterion.

Table 6

Detection Rates for an Item With True Differential Item Functioning of 1.62

Sample sizes	Refined	Unrefined
$n_R = 200$, $n_F = 50$	8.4	9.2
$n_R = n_F = 500$	6.0	15.4

Note. The test had 20 items with four DIF items, balanced DIF, and N (-1,1) focal group ability distribution. DIF = differential item functioning.

Recommendations—Criterion Refinement

Although it is often assumed that refinement always provides superior results, the actual situation proves to be more complex. If previous research or theoretical considerations suggest that DIF is likely to be balanced, then the unrefined approach 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 on average, it is only slightly disadvantageous in balanced conditions, whereas the unrefined method tends to have substantially lower detection rates in unbalanced conditions.

Overall Summary and Discussion

DIF analysis is an essential element in the evaluation of the fairness and validity of educational tests. It is important that these analyses produce accurate and useful results. This project reviewed three aspects of ETS DIF procedures, focusing on the case of dichotomous items: the nature and stringency of the statistical rules used to flag items, the minimum sample

size requirements that are currently in place for DIF analysis, and the efficacy of criterion refinement.

One conclusion of the review was that the ETS C rule often displays low DIF detection rates even when samples are large. A review of some possible alternative rules suggests that higher detection rates can be achieved without incurring excessive Type I error. Therefore, a reconsideration of the current flagging rules is recommended. A review of this kind should start with a reevaluation of the minimal size of DIF that is important to detect and the degree of false positives that can be tolerated. Because the utility of the B category is questionable, it may be advantageous to explore the possibility of establishing a two-category DIF classification system instead of the current three-category system. It is worth noting that determination of the smallest DIF that is important to detect need not be in terms of the delta metric, which is unlikely to be well understood outside ETS. The odds-ratio metric, discussed earlier in the paper, and the proportion-correct metric, as embodied in the *STD P-DIF* statistic of Dorans and Kulick (1986), are candidates for consideration.

In his review of this paper, Neil Dorans made the further suggestion that future DIF flagging criteria could perhaps take into account the potential impact on test-takers of the presence of DIF in the situation at hand. Thus, the flagging criteria could take into account the number of items on the test, the way the test is scored, and the way the scores are used.

A second finding of this review, which is related to the stringency of flagging criteria, is the conclusion that an improvement of the flagging rules could allow minimum sample size requirements to be relaxed. The determination of minimum sample sizes can also be guided by the decisions that are made about the minimal size of DIF that is important to detect and the amount of Type I error that is tolerable. In addition, there appear to be several satisfactory ways of aggregating DIF information across multiple administrations or administration windows (the simplest of which is to combine the data and perform the standard analyses). Therefore, it might be useful to relax the guidelines for doing so rather than simply exempting programs from performing DIF analyses for small groups.

A third conclusion is that refinement of the matching criterion is helpful when DIF is unbalanced (i.e., primarily in one direction) but can be detrimental when DIF is balanced. If nothing is known in advance about the likely pattern of DIF, the MH procedure with refinement is advisable since its overall accuracy rate is higher than that of the unrefined procedure.

In addition to the analyses of flagging rules, minimum sample size, and criterion refinement, this DIF review also included a web-based survey of individuals involved in DIF analysis, review, or research at ETS. The goal of the survey, which was conducted in May 2011, was to help identify DIF issues perceived as being "most in need of further clarification, examination or reconsideration at ETS." The respondents were asked to select these issues from a checklist of 17 possible responses.

The items that were checked by at least 10 respondents were the following (in order of popularity): DIF analysis procedures for small samples, minimum sample size requirements, DIF analysis procedures for polytomous items, DIF analysis procedures for complex performance tasks, DIF analysis procedures for computerized adaptive tests, inclusion/exclusion of non-U.S. citizens and those for whom English is not the best language, and interpretability of DIF results by staff and review panels. The two most-endorsed items, then, concern sample size, the next three involve DIF analyses for specialized types of assessment, the sixth involves group composition, and the last is the key issue of interpretability. It is hoped that the findings of the DIF survey will be helpful in designing DIF analyses and DIF review processes and in crafting agendas for future research.

Sample size issues were considered in the present study, and two other issues identified in the survey, DIF procedures for polytomous items and DIF procedures for performance tasks, are being addressed in a related ETS project: Tim Moses, Jinghua Liu, Adele Tan, Weiling Deng, and Neil Dorans have been conducting research on DIF analyses of constructed-response items as they are conducted at ETS (Tim Moses, personal communication, July 22, 2011). This project has been evaluating the various ways of defining matching variables for mixed format tests. These matching variables can be based on scores on the constructed-response items, scores on the multiple-choice items, or a sum or bivariate combination of the constructed-response and multiple-choice scores. The Moses et al. project has also considered the issues of inclusion of the studied item in the matching variable and the use of observed-score versus model-based matching variables.

Other technical issues that could be considered in a more comprehensive review of ETS DIF procedures are the definition of appropriate groups for DIF analysis, the possibility of conducting multiple-group rather than pairwise DIF analyses, and the optimal DIF analysis procedures for formula-scored tests, tests scored using item response theory scales, and

computerized adaptive tests. A further review could also comprise a consideration of policy issues, such as the rules for establishing DIF committees and determining what information to present them.

Much has changed since ETS began implementing operational DIF procedures 25 years ago. New forms of assessment have been developed, test scoring has become more sophisticated, and definitions of racial and ethnic categories have been modified. The time is ripe for a reconsideration of ETS DIF policies and procedures.

References

Camilli, G., & Smith, J. K. (1990). Comparison of the Mantel-Haenszel test with a randomized and a jackknife test for detecting biased items. *Journal of Educational Statistics, 15*, 53–67.

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.

Davey, T., & Wendler, C. (2001, April 3). *DIF best practices in statistical analysis* [ETS internal memorandum].

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

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

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*, 309–319.

Educational Testing Service. (1988, November). *Procedures for use of differential item difficulty statistics in test development* [ETS internal memorandum].

Fidalgo, A. A., Ferreres, D., & Muñiz, J. E. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement, 64*, 925–936

Fidalgo, A. M., Hashimoto, K., Bartram, D., & Muñiz, J. (2007). Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. *Journal of Experimental Education, 75*, 293–314.

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. (1987a, January 27). *Expansion and comments on Marco's rational approach to flagging items for DIF* [ETS internal memorandum].

Holland, P. W. (1987b, February 11). *More on rational approach item flagging* [ETS internal memorandum].

Holland, P. W. (2004, February 9). *Comments on the definitions of A, B, and C items in DIF* [ETS internal memorandum].

Holland, P. W., & Thayer, D. T. (1985). *An alternate definition of the ETS delta scale of item difficulty* (ETS Program Statistics Research Technical Report No. 85-64). Princeton, NJ: ETS.

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, CA. Retrieved from the ERIC database. (ED272577).

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

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.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the *MH D-DIF* statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Hillsdale, NJ: Erlbaum.

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, N.J: Erlbaum.

Lu, S.-M., & Dunbar, S. B. (1996, April). *Assessing the accuracy of the Mantel-Haenszel DIF statistic using bootstrap method.* Presented at the annual meeting of the American Educational Research Association, New York.

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.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32*, 302–316.

Petersen, N. S. (1987, September 25). *DIF procedures for use in statistical analysis* [ETS internal memorandum].

Sinharay, S., Dorans, N. J., Grant, M. C., & Blew, E. O. (2009). Using past data to enhance small sample DIF estimation: A Bayesian approach. *Journal of Educational and Behavioral Statistics, 34*, 74–96.

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.

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

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

Zwick, R., & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel DIF analysis to a computerized adaptive test. *Applied Psychological Measurement*, *26*, 57–76.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel- Haenszel DIF analysis. *Journal of Educational Measurement*, *36*, 1–28.

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. (in press). Improving Mantel-Haenszel DIF estimation through Bayesian updating. *Journal of Educational and Behavioral Statistics*.

Zwick, R., Ye, L., & Isham, S. (2012, April). *Investigation of the efficacy of DIF refinement procedures.* Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.

Notes

[1] The standardized proportion difference (*STD P-DIF*) statistic of Dorans and Kulick (1986) is used descriptively to supplement the Mantel-Haenszel results. In addition, standardized distractor analysis (Dorans, Schmitt, & Bleistein, 1992) is used to examine the answer choices made by the reference and focal groups, conditional on the matching criterion.

[2] In recent DIF analyses of actual data with approximately 2,000 members per group, roughly two-thirds of the items had standard errors small enough to render irrelevant the statistical significance criterion in Equation 4. About 90% of the items had standard errors small enough to render irrelevant the criterion in Equation 8.

[3] Item responses were simulated using the three-parameter logistic (3PL) model, and DIF was induced by introducing a difference between the reference and focal group difficulties. Translating the DIF into the MH metric to obtain the true DIF values was accomplished using a formula given in Zwick et al. (in press). For theoretical reasons, the *MH DIF* procedure is expected to perform optimally under the Rasch model and less well under the 3PL (see Holland & Thayer, 1988; Zwick, 1990). However, the 3PL model produces data much more similar to those that result from actual administration of multiple-choice tests and was therefore used in this study.

[4] If a more conventional definition of Type I error had been used, Type I error rates would be lower than the tabled rates.

[5] Longford, Holland, and Thayer (1993, p. 182) proposed similar DIF estimates, although they did not use a Bayesian framework. Their approach was based on a random effects model for DIF. Maximum likelihood estimates of the model parameters were obtained through an iterative procedure.

[6] It is useful to note as well that continuity corrections are intended to ensure conservative inferences (Shelby Haberman, personal communication, November 15, 2011).

[7] See Zwick et al. (in press) for a discussion of the reasons for biases in MH D-DIF statistics in unbalanced conditions. See Table 4 in the present paper for an example of bias results.