Research Report

Using Past Data to Enhance Small-Sample DIF Estimation: A Bayesian Approach

Sandip Sinharay

Neil J. Dorans

Mary C. Grant

Edwin O. Blew

Colleen M. Knorr

Research & Development

May 2006
RR-06-09

Using Past Data to Enhance Small-Sample DIF Estimation: A Bayesian Approach

Sandip Sinharay, Neil J. Dorans, Mary C. Grant, Edwin O. Blew, and Colleen M. Knorr
ETS, Princeton, NJ

May 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

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

Copyright © 2006 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, GRADUATE RECORD EXAMINATIONS, GRE, and PPST are registered trademarks of Educational Testing Service (ETS). SAT is a registered trademark of the College Board.

Abstract

The application of the Mantel-Haenszel test statistic (and other popular DIF-detection methods) to determine DIF requires large samples, but test administrators often need to detect DIF with small samples. There is no universally agreed upon statistical approach for performing DIF analysis with small samples; hence there is substantial scope of further work on the problem. One advantage of a Bayesian approach over a frequentist approach is that the former can incorporate, in the form of a prior distribution, existing information on the inference problem at hand; a prior distribution often leads to improved estimation, especially for small samples. Further, for any operational test, a huge volume of past data is available, and for any item appearing in a present test, there is a high chance that a number of similar items have appeared on past operational administrations of the test. Therefore, ideally, it will be possible to use that past information as a prior distribution in a Bayesian DIF analysis. This paper discusses how to perform such an analysis. The suggested Bayesian DIF analysis method is shown to be an improvement over the existing methods in a realistic simulation study.

Key words: Empirical Bayes, loss function, Mantel-Haenszel statistic, prior distribution

Acknowledgments

This project would not have been possible without the sincere help of Kevin Larkin, Shaloo Gupta, and Stacie Rupp with the data analyzed in this study. The authors also thank Gautam Puhan, Charles Lewis, Robert Mislevy, Dan Eignor, Paul Holland, John Donoghue, Mark Gierl, Rebecca Zwick, Hal Stern, and Shelby Haberman for useful advice and Amanda McBride for editorial help.

1 Introduction

Differential item functioning (DIF) refers to a difference in test-item functioning between two comparable groups of examinees, that is, groups that are matched with respect to the construct measured by the test. Holland (1985) suggested the Mantel-Haenszel (MH) test statistic, arguably the most popular statistic for the purpose, to detect whether an item shows DIF. The application of the MH statistic requires large samples. Clauser and Mazor (1998) commented that "samples of 200 to 250 per group have been consistently shown to be suitable for use with the Mantel-Haenszel statistic" (p. 37). In practice, test administrators often face the dilemma of detecting DIF with small samples. Further, even for tests with several thousand examinees, there may be only a few examinees belonging to a group of interest, such as Native Americans. However, there is no universally agreed upon statistical approach for performing DIF analysis with small samples, and hence there is substantial scope for further work on the problem.

One advantage of the Bayesian statistical methods over frequentist methods is that the former can incorporate, in the form of a prior distribution, existing information on the inference problem at hand, leading to improved estimation, especially for small samples for which the posterior distribution is sensitive to the choice of prior distribution. Further, for most operational tests, a huge volume of past data is available, and for any item appearing in a current test, a number of similar items are often found to have appeared in past operational administrations of the test. Conceptually, it should be possible to incorporate that past information into a prior distribution in a Bayesian DIF analysis.

Bayesian methods have been applied to DIF analysis. Zwick, Thayer, and Lewis (1999, 2000) and Zwick and Thayer (2002) applied empirical Bayes (EB) methods to DIF analysis and found the EB estimate of DIF to be an improvement over the operationally used MH D-DIF statistic (Holland & Thayer, 1988), especially for small samples. The EB method estimates the prior mean and variance from the current data and uses the same prior information for all the items. However, past information may suggest different prior distributions for different item types that can be used in a DIF analysis using a full Bayesian (FB) approach. No rigorous works exist on FB methods for DIF analysis, except a small study by Lewis and Thayer (2000), who did not use past data to form their prior distributions, but, rather, used noninformative prior distributions and found the FB method offered no improvement over the EB method.

Motivated by the above, this paper suggests an FB DIF estimation method that uses past

information in an attempt to improve DIF analysis. The past information is quantified into prior distributions, one for each item type, which are then used in the FB estimation technique.

One can employ other approaches for small-sample DIF estimation (e.g., exact Mantel-Haenszel tests suggested by Parshall & Miller, 1995, and Meyer, Huynh, & Seaman, 2004),[1] but those are not discussed in this paper.

Section 2 provides some background—it introduces the ETS operational DIF-detection method, discusses the EB approach to DIF estimation, and explains the motivation of the current work. Section 3 discusses the suggested FB approach. Section 4 and 5 provide the study design and results, respectively. Section 6 provides discussion and conclusions.

2 Background

2.1 The ETS Operational DIF-Detection Method (the Mantel-Haenszel Statistic)

Holland (1985) suggested the MH test statistic for studying DIF in the context of item response data. Suppose one is interested in examining if a given item j shows DIF for a focal group, F, (which is of primary interest) and a reference group, R. In an application of the Mantel-Haenszel test, the examinees are divided into K-matching groups; the groups are typically formed based on examinees' total raw scores. For an item, the data from the k-th matched group of reference and focal group members are arranged as a 2×2 table, as shown in Table 1. The MH

Table 1.

Mantel-Haenszel Statistic for k-th Matched Groups

	Right on an item	Wrong on an item	Total
Reference	A_k	B_k	n_{Rk}
Focal	C_k	D_k	n_{Fk}
Total	R_k	W_k	n_{+k}

odds ratio estimate (Mantel & Haenszel, 1959) for an item, which compares the two total groups in terms of their odds of answering the item correctly conditional on the proficiency measure, is then given by

$$\hat{\alpha}_{\text{MH}} = \frac{\sum_k A_k D_k / n_{+k}}{\sum_k B_k C_k / n_{+k}}. \tag{1}$$

The MH index of DIF, *MH D-DIF* (also denoted as Δ_{MH}), suggested by Holland and Thayer (1988), is given by

$$MHD - DIF \equiv -2.35 \times \log_e(\hat{\alpha}_{\mathrm{MH}}). \tag{2}$$

The above transformation places MH D-DIF on the ETS delta scale of item difficulty (Holland & Thayer, 1985). By definition, MH D-DIF is defined so as to be negative when the item is more difficult for members of the focal group than it is for the comparable members of the reference group. Phillips and Holland (1987) estimated the variance of $\log_e(\hat{\alpha}_{\mathrm{MH}})$ by

$$\frac{1}{2\left(\sum_k A_k D_k / n_{+k}\right)^2} \sum_k \left(A_k D_k + \hat{\alpha}_{\mathrm{MH}} B_k C_k\right)\left(A_k + D_k + \hat{\alpha}_{\mathrm{MH}}(B_k + C_k)\right)/(n_{+k})^2. \tag{3}$$

ETS has a system of categorizing the extent of DIF based on both the magnitude of the MH D-DIF index and the statistical significance of the results (see, e.g., Dorans & Holland, 1993). According to this scheme, an item has a DIF classification of C, representing moderate to large DIF, if the absolute value of MH D-DIF is at least 1.5 and is significantly greater than 1 (at a 5% significance level). An item has a DIF classification of A, representing negligible DIF, if either the absolute value of MH D-DIF is less than 1 or if the MH D-DIF value is not significantly different from 0. Items that cannot be classified as A or C are considered to have slight to moderate DIF and have a DIF classification of B. Items that have been classified as C are subjected to thorough scrutiny and are usually eliminated from tests. Because it is often important to distinguish between negative DIF (DIF against the focal group) and positive DIF (DIF favoring the focal group), the above rules result in five DIF classifications or categories: C-, B-, A, B+, C+, where the + and - signify positive and negative DIF, respectively.

2.2 Review of the Results Using an Empirical Bayes Approach to DIF

The description of the empirical Bayes approach. Zwick, Thayer, and Lewis (1999) argued that the operational DIF classification system conveys the notion that an item's DIF category is deterministic; they suggested using a Bayesian approach to estimate the probabilities that the true DIF for an item falls into the A, B, or C categories (called the *true DIF method*) or to estimate the probabilities that an item would be classified as A, B, or C in future administrations (called the *future DIF method*).

Zwick et al. (1999) assumed the MH D-DIF statistic, defined in (2) and denoted henceforth as Δ_i (for item i), followed a normal distribution given by

$$\Delta_i \sim \mathcal{N}(\theta_i, \sigma_i^2) \tag{4}$$

conditional on $\theta_i = E(\Delta_i)$, where θ_i represents the unknown DIF parameter value corresponding to Δ_i, and σ_i^2 denotes the sampling variance of the MH statistic. The variance σ_i^2 is treated as known and set equal to the observed estimate of the squared standard error $S_i^2 = SE^2(\Delta_i)$. The prior distribution assumed was

$$\theta_i \sim \mathcal{N}(\mu, \tau^2), \tag{5}$$

where μ and τ^2 are the overall mean and variance of the θ_is for the test under consideration. Equations 4 and 5 imply that the posterior distribution of θ_i given Δ_i, μ, and τ^2 is given by

$$f(\theta_i|\Delta_i, \mu, \tau^2) \propto f(\Delta_i|\theta_i) f(\theta_i|\mu, \tau^2) \equiv \mathcal{N}\left(W_i\Delta_i + (1 - W_i)\mu, W_i S_i^2\right), \tag{6}$$

where

$$W_i = \frac{\tau^2}{S_i^2 + \tau^2}. \tag{7}$$

The above equation shows that the posterior mean of θ_i is a *shrinkage estimator*. As the value of S_i^2 becomes large (which means considerable error in estimating θ_i), W_i approaches 0, and the EB estimation procedure shrinks the posterior mean considerably toward the prior mean μ. On the other hand, as S_i^2 approaches 0 (which means small error in estimating θ_i), W_i approaches 1, and the posterior mean approaches the observed value Δ_i.

Zwick et al. (1999) then adopted an empirical Bayes (EB) approach (Braun, 1989; Robbins, 1955) to estimate the model, which involved estimation of μ and τ^2 from the available data. Camilli and Penfield (1997) and Zwick et al. showed that EB estimates are given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\Delta_i, \quad \hat{\tau^2} = \widehat{\mathrm{Var}(\Delta_i)} - \frac{1}{n}\sum_{i=1}^{n} S_i^2, \tag{8}$$

where $\widehat{\mathrm{Var}(\Delta_i)} = \frac{1}{n-1}\sum_{i=1}^{n}(\Delta_i - \hat{\mu})^2$, the across-item variance of the MH D-DIF estimates. The above estimates of μ and τ^2 are then plugged in Equation 6 to obtain the (estimated) posterior distribution of θ_i.

Zwick, Thayer, and Lewis (2000) investigated a DIF flagging method based on the loss function (where *keep* means *failure to flag*)

$$L(\text{keep}) = c\theta_i^2, \quad L(\text{flag}) = k\cdot \tag{9}$$

Suppose the *indifference point*, that is, the value of θ_i for which $L(\text{keep}) = L(\text{flag})$, is θ_*. This means that if the true DIF of the item, expressed in the MH metric, is equal to θ_* in absolute value, we are indifferent as to whether the item is kept or flagged, which implies $c\theta_*^2 = k$. If the true DIF is smaller than θ_*, we want to avoid flagging the item, whereas if the true DIF exceeds θ_*, we would want to flag the item. The decision rule that minimizes the posterior expected loss is

$$\text{Flag if } cE(\theta_i^2|\Delta_i) \geq k, \text{ i.e., if } cE(\theta_i^2|\Delta_i) \geq c\theta_*^2. \text{ i.e., if } E(\theta_i^2|\Delta_i) \geq \theta_*^2\cdot \tag{10}$$

The average loss related with the above rule is

$$L(\text{keep})P(\text{keep}) + L(\text{flag})P(\text{flag}). \tag{11}$$

Zwick et al. used $\theta_* = 1$ (which leads to $c = k$) according to the recommendation by Holland (1987a, 1987b). In terms of the ABC magnitude criteria, this is equivalent to a rule that flags B and C items. That is, in very large samples where statistical significance is irrelevant, we would expect the loss function with an indifference point of 1 to show essentially the same results as a rule that flags both B and C items. *Summarization of the results obtained using the EB method.* Zwick et al. (1999) performed a number of simulations and real data analyses, comparing their EB approach to the operational MH D-DIF approach. For small sample sizes, the results often differed. For example, for an item in the verbal section of the 1990 Graduate Record Examination® (GRE®), the operational DIF category was C for comparing White (sample size 8,736) versus Asian-American (sample size 220) students. However, the true DIF method estimated the chance of the item belonging to the C category to be 4% and the chance of the item belonging to the A category to be 65%. In the simulation studies, the EB point estimates (posterior means) had considerably less root mean squared error (RMSE),

$$\frac{1}{\text{Nrep}} \sum_{j=1}^{\text{Nrep}} (E_j - \text{True DIF})^2,$$

where Nrep is the number of replications, and E_j is the estimate in the j-th replication, than the standard MH D-DIF statistics. For small sample sizes, the superiority of the EB method was

greatly increased. For example, in a case with 200 reference group members and 50 focal group members described in Table 7 of Zwick et al. (1999), the median RMSE was 0.65 for the EB estimate against 0.97 for the MH D-DIF statistic (the EB estimate had less RMSE for 33 items out of a total of 36 items). The smaller RMSE values for the EB estimates found in Zwick et al. were consistent with well-established theory. According to the Stein effect (James & Stein, 1961), estimates can be improved by using information from all coordinates (in this case the MH D-DIF values for all items) in estimating each coordinate. Such estimates have smaller mean squared error than their non-Bayesian counterparts. Though EB estimators are not unbiased (whereas MH estimates are), lower RMSE more than makes up for the bias.

Zwick et al. (1999) further considered repeated administrations of the same test form for the SAT® examinations in 1989 and 1991 and GRE examinations in 1991 and 1992. For both the SAT and GRE examinations, point estimates of DIF parameters computed (both EB and MH D-DIF) from Time 1 were compared to the Time 2 MH D-DIF statistics. Zwick et al. found the RMSEs for the EB estimates to be smaller than those for the Time 1 MH D-DIF statistics (except for one case). The advantage of the EB method (with regard to point estimates) was more when sample sizes were small (50 for the focal group and 200 for the reference group); for larger sample sizes, the prior distribution has little influence and the EB estimate is close to the MH estimate. The advantage was more for the mathematics or quantitative items than for verbal items.

Zwick et al. (2000) compared their loss-function-based EB method to two possible flagging approaches based on the MH D-DIF statistic: (a) B rule (which flags both B and C items; this should be comparable to the EB rule) and (b) C rule (which flags only C items). In simulations under reasonable assumptions about the relative seriousness of Type I and Type II errors, the authors found the loss-function-based DIF-detection rule to perform better overall than the B rule and the C rule, especially for small samples. However, compared to the B rule and the C rule, the loss-function-based DIF-detection rule was more likely to correctly identify items with DIF and had a comparatively higher Type I error rate (p. 244)—that is, the loss-function-based DIF-detection rule was more likely to incorrectly flag items that did not have DIF (p. 238).

Zwick and Thayer (2002) performed a simulation study to investigate the applicability of the method developed by Zwick et al. (1999, 2000)to computer-adaptive test (CAT) data. Results showed the performance of the EB DIF approach to be quite promising, even in extremely small

samples. In particular, the EB procedure was found to achieve roughly the same degree of stability for samples averaging 117 and 40 members in the two examinee groups as did the ordinary MH statistic for samples averaging 240 in each of the two groups (the median RMSE achieved for EB with samples averaging 117 and 40 members was the same as the median RMSE achieved for MH for sample sizes averaging 240). Overall, the EB estimates tended to be closer to their target values than did the ordinary MH statistics in terms of RMSE; the EB estimates were also more highly correlated with the target values than were the MH estimates.

Lewis and Thayer (2000) applied the EB method and an FB method (with noninformative prior (i.e., not using any past data) to data from an adaptive licensure test. Using two sets of items that appeared both in a pretest (small sample size) and an operational test (large sample size), they used pretest information to predict operational performance of the different methods in two ways: (a) they computed pretest-based point estimates to predict operational values of the MH D-DIF statistics, and (b) they computed pretest-based classifications to predict operational classifications. The research found both the EB and FB estimates to perform better than the MH D-DIF statistic in terms of RMSE. The researchers also found the EB method to perform slightly better than the FB estimate. However, the results were opposite for classification; the MH D-DIF method performed better than both the Bayesian methods with respect to classification. Lewis and Thayer argued that using the two Bayesian estimates provided a better combination of prediction and classification than would using pretest MH D-DIF statistics alone. The researchers also recommended that, to produce a test that is as fair as possible for all test takers, all sample-size restrictions currently in place be dropped for the licensure test they dealt with, which had a surplus of items available.

2.3 Existing Literature on the Use of Past Information in Bayesian Methods

An advantage of the Bayesian methods over the frequentist methods is that the former can incorporate prior information about the problem into the statistical model in the form of a prior distribution. For an inference problem with small sample size, where the prior distribution affects the posterior distribution substantially, the advantage can be practically significant. In educational testing, this may have major implications. Often, as in the problem of small-sample DIF detection (which is the focus of this paper), the size of the current sample is small, but there are usually a large amount of past data easily available (most of the tests administered over

the years do not change too much over time). Therefore, ideally, if prior distributions can be formed using past data, there is a scope of improvement in the quality of statistical estimation for applications with small sample sizes. To form a prior distribution, one has to find a way to quantify the information from the past data in an optimum way into a prior distribution.

Gelman, Carlin, Stern, and Rubin (2003, p. 260) had an example where past data were used to elicit a prior distribution for a statistical model applied to pharmacokinetics. In an application of Bayesian statistics in medical screening, Johnson and Gastwirth (1991, p. 435) elicited a prior distribution from past data; specifically, they treated the posterior distribution of a parameter from the analysis of Canadian people as the prior distribution for analysis of British people.

In psychometrics, Novick and Jackson (1974, chapters 6-9 and 7-12) had a number of examples where prior distributions were constructed using past data, in the context of a beta-binomial model (where the response variable has a binomial distribution whose success probability follows a beta prior distribution) and normal model.

To apply the three-parameter logistic (3PL) model to 1987 American College of Testing (ACT) mathematics test data, Tsutakawa (1992) formulated a prior distribution for the model parameters from 1981 ACT mathematics test data.

Recently, Swaminathan, Hambleton, Sireci, Xing, and Rizavi (2003) found in a simulation study that estimation of item response theory (IRT) model parameters can be improved considerably by incorporating judgmental data on item difficulty into the prior distribution for the difficulty parameter. Other than these few studies, there is a lack of work on constructing prior distributions from past data in psychometrics, even though Bayesian statistics has become extremely popular in the field (see, e.g., Sinharay, 2005) and the sample size in such applications is often not large.

2.4 Previous Research on Item Characteristics Affecting DIF

Schmitt and colleagues, through a number of well-planned studies, revealed a number of factors that contribute to DIF. Schmitt (1985, 1988) found, among other things, that true cognates tended to have positive DIF for Hispanic examinees taking the SAT examination. Schmitt and Bleistein (1987) reported that Black students did not complete SAT-V sections at the same rate as White students with comparable SAT-V scores and that this differential speededness effect appeared to account for much of the negative DIF for Blacks on SAT verbal analogy items.

Dorans, Schmitt, and Curley (1988) found a similar phenomenon (also see Schmitt, Dorans, & Holland, 1993) in a special confirmatory experiment. Schmitt and Dorans (1990) analyzed SAT data to report findings such as (a) an item whose content is of special interest shows positive DIF for the relevant ethnic group, and (b) homographs tend to show negative DIF for Asian American, Hispanic, and Black students. Schmitt, Curley, Bleistein, and Dorans (1988) reported the results from a randomized DIF study where specially constructed items, developed to test specific hypotheses, were administered under conditions that permitted appropriate statistical analyses to assess the efficacy of the hypotheses. The most convincing support was found for the hypothesis that the true cognates show positive DIF for Hispanic examinees.

The research of Lawrence and Curley (1989) and Lawrence, Curley, and McHale (1988) on SAT reading passages showed that content related to technical aspects of science (as opposed to the history or philosophy of science) appeared to be more difficult for women than for a matched group of men.

Gallagher et al. (2000) proposed a taxonomy of content and cognitive characteristics to account for gender differences in mathematics. However, Gierl, Bisanz, Bisanz, and Boughton (2003) found little support for the taxonomy in a carefully designed study of factors leading to DIF in a curriculum-based mathematics achievement test for Canadian ninth-graders; they commented (p. 299) that the taxonomy may not be entirely adequate. The only characteristic mentioned in Gallagher et al. that was supported by the results in Gierl et al. is that items that require significant spatial processing show negative DIF for women. O'Neill and McPeek (1993) and Camilli and Shepard (1994) also reported several factors causing DIF in a variety of situations.

3 A Full Bayesian Approach Using Past Information

Lord (1986) argued, in the context of IRT parameter estimation, the following:

> When approximately parallel test forms are administered year after year to similar population of examinees, it becomes possible to deduce appropriate prior distributions for the item and the ability parameters from past results. In such a situation, Bayesian procedures should certainly yield better parameter estimates than maximum likelihood, since Bayesian procedures make use of more information. (p. 158)

The argument applies for DIF estimation as well, as is clear from the following comment of Zwick et al. (1999), "It is anticipated that in operational applications of the EB methods to test data, the results of previous DIF analyses of the same test could also be used to inform the selection of appropriate values for μ and τ^2" (p. 7).

Our work basically implements the above idea. We not only restrict ourselves to one μ and τ^2 as mentioned by Zwick et al. (1999), but we also allow μ and τ^2 to vary over the different types of items. The above discussions on the use of past information in Bayesian methods and previous research on item characteristics affecting DIF motivated our work.

The EB method of Zwick et al. (1999), while itself an improvement over the current operational MH D-DIF method, uses the same prior information for all the items while past information may suggest different prior distributions for different item types that can be used in a DIF analysis using a full Bayesian (FB) approach.

In our analysis, similar to Zwick et al. (1999), the basic statistical model is

$$\Delta_{iT} \sim \mathcal{N}(\theta_{iT}, \sigma^2_{iT}), \tag{12}$$

where Δ_{iT} denotes the MH D-DIF statistic for item i that is of a specific item type T. For example, T could denote items with content related to science. As in Zwick et al., the σ^2_{iT}s are assumed known and set equal to $S^2_{iT} = SE^2(\Delta_{iT})$, the observed estimate of the squared standard error. We assume the prior distribution to be

$$\theta_{iT} \sim \mathcal{N}(\mu_T, \tau^2_T), \tag{13}$$

that is, unlike in Zwick et al. (1999), the true DIF parameter of an item is assumed to follow a normal distribution with mean and variance specific to the type of that item. The mean and variance of the MH D-DIF estimates of past items of type T are used as estimates of μ_Ts and τ^2_Ts, respectively, and are denoted as $\widehat{\mu_T}$ and $\widehat{\tau^2_T}$. Then, the posterior distribution of θ_{iT} is given by

$$f(\theta_{iT}|\Delta_{iT}) \equiv \mathcal{N}\left(\widetilde{W_{iT}}\Delta_{iT} + (1 - \widetilde{W_{iT}})\widehat{\mu_T}, \widetilde{W_{iT}}S^2_{iT}\right), \tag{14}$$

where

$$\widetilde{W_{iT}} = \frac{\widehat{\tau^2_T}}{S^2_{iT} + \widehat{\tau^2_T}}. \tag{15}$$

This approach has the potential to provide more accurate results in DIF analysis with small sample sizes than an empirical Bayes method or a frequentist method. Earlier discussions in

this paper showed several experts believed that some items are more likely to show DIF than others and that information quantified in the prior distribution should provide an advantage to the FB method. From the description of the approach, it is clear that the FB approach requires more information than the traditional MH approach or the EB approach. For assessments with well-maintained records, however, obtaining the extra information should not be costly. It is also clear that computation using the approach requires little extra time compared to the MH or the EB approach.

4 Study Design

To compare the performance of the FB estimates to the MH and EB estimates, we performed a realistic simulation study that used operational test data. We used data from 12 administrations (between 2001 and 2004) of the Pre-Professional Skills Test (PPST®) on reading. These test forms were administered to between 7,000 and 19,000 examinees. The prior distributions were constructed from the 10 least recent administrations (considered as *past data*). Then, to study the performance of the DIF estimation methods, we iterated the following two steps on data from each of the two remaining, most recent administrations (which were treated as *current data*) 1,000 times:

1. Choose a random subsample from the data. Two different sample sizes were used. To study the results for a small-sample size condition, we set the size of the smaller group (male examinees for male-female DIF and Black examinees for White-Black DIF) as 50. As in the operational analysis, the focal group for male-female DIF analysis is the female group and the focal group for White-Black DIF analysis is the Black group throughout this paper (note that in the male-female DIF, the smaller group of male examinees is the reference group). To study the results for a larger sample-size condition, we set the size of the smaller group as 300. The size of the other group was chosen to maintain the same ratio of focal group size and reference group size as in the full sample. For example, the number of males and females in Data Set 1 for the PPST reading test were 2,736 and 10,203, respectively. To maintain the ratio of male-to-female at 2736:10203 and to keep the smaller group size (that of male examinees) at 50 and 300 for the focal and the reference groups, respectively, we needed 186 and 1,120 female examinees in the subsamples.

2. Compute the MH D-DIF statistic, EB DIF estimate, and FB DIF estimate for each item from the small subsample. Also obtain the DIF classification provided by each method for each item from the small subsample. For the MH statistic, we considered two flagging approaches as in Zwick et al. (2000): a B rule (that flags both B and C items) and a C rule (that flags only C items). While the B rule is comparable to the EB and FB method directly, the C rule is mostly used operationally. For the EB or FB method, obtaining the DIF classification required use of the decision rule given in Equation 10.

The above provides, for each item, 1,000 point estimates and 1,000 DIF classifications using each approach. We then compared the point estimates to the *true DIF*, which is the value of MH D-DIF from the full sample, using RMSE and squared bias measures, where, for an item and an estimation method,

$$\text{Bias} = \frac{1}{1000}\sum_{j=1}^{1000} \text{Estimated DIF from subsample } j - \text{True DIF},$$

$$\text{RMSE} = \sqrt{\frac{1}{1000}\sum_{j=1}^{1000} (\text{Estimated DIF from subsample } j - \text{True DIF})^2}.$$

We also determined the true DIF classifications for the items. As in Zwick et al. (2000, p. 238), because the indifference point is 1, true DIF classification of items with true DIF values less than one in absolute value is *keep* and the true DIF classification of any other item is *flag*. We compared the DIF classifications from the subsamples to the true DIF classifications to obtain average percent correct decisions for keep items, flag items, and overall for each of the methods. Also, the average loss for keep items, flag items, and overall are obtained using the average loss function given by Equation 11. Here, as the true DIF values θ_i are assumed known, $L(\text{keep}) = c\theta_i^2$ can be easily computed; $L(\text{flag}) = k$ is set to 1 to fix the scale, and $P(\text{keep})$ and $P(\text{flag})$ are estimated from the simulation results.

The nature of the simulation study used in this paper is rare and hence needs special mention. The simulations use subsamples of operational data instead of data generated from an IRT model, which, as noted in Wainer and Thissen (1987), often have an uncertain relationship to reality. Thus, these simulations provide realistic results.

5 Results for PPST Reading Assessment

5.1 Analysis of Past Data

For the PPST reading assessment (PPSTR), test developers classify each item into one of two D1 classifications according to the item type. Each item is also classified into one of two D2 classifications based on the item's content. The D1 and D2 classifications for the PPST assessments are designations for various types and levels of item classification. These, along with a number of other classifications, are used by the test developers to help in assembling test forms that conform to specifications and constraints. For example, in PPSTR, the two D1 classifications are *literary comprehension* and *critical and inferential comprehension.* For security purposes, no further details of these classifications are included.

Figures 1 and 2, respectively, show box plots for the distributions of the male-female and White-Black MH D-DIF statistics for the different D1 and D2 classifications for 400 items from 10 past forms of the PPSTR assessment.

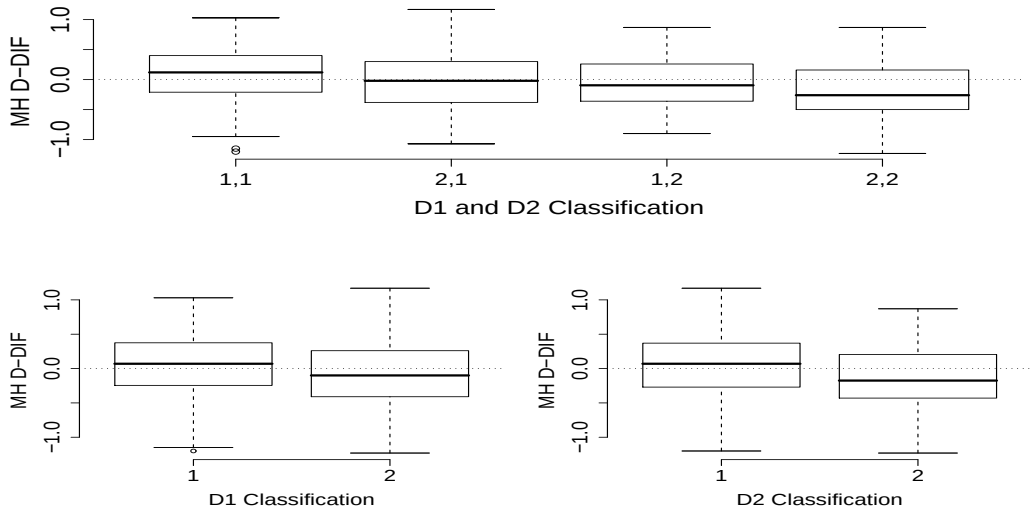

***Figure* 1. Box plots for the MH D-DIF statistics for PPSTR for male-female DIF.**

Note. The dotted line corresponds to an MH D-DIF value of 0.

The figures show that although the average of the MH D-DIF statistics is close to 0 for any item type, there is some variation among the classifications with respect to the corresponding distributions of MH D-DIF statistics. For example, items with the second D2 classification more often have negative MH D-DIF statistics for White-Black DIF analysis.

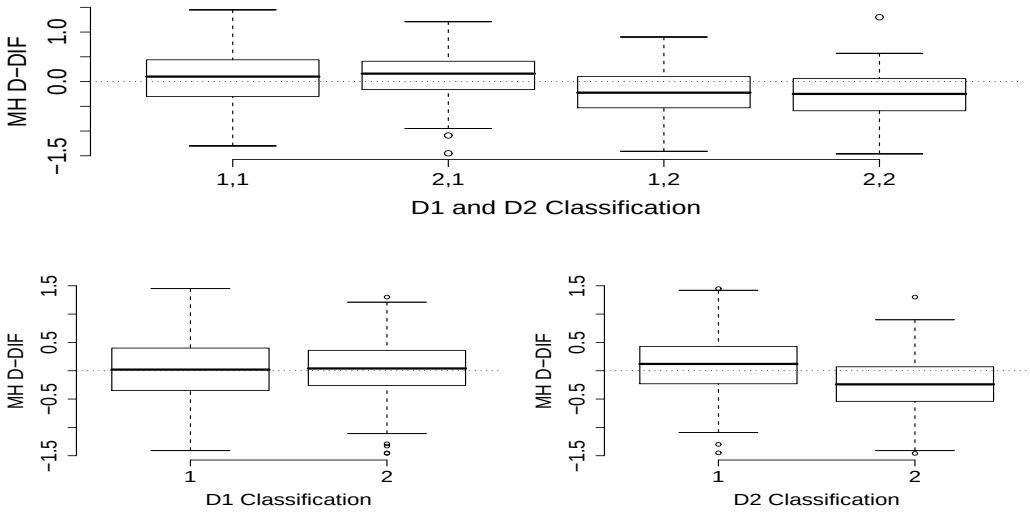

***Figure* 2. Box plots for the MH D-DIF statistics for PPSTR for White-Black DIF.**

Note. The dotted line corresponds to an MH D-DIF value of 0.

For male-female DIF, none of the 400 past items have C DIF and only 10 have B DIF. For White-Black DIF, none of the 400 past items have C DIF and only 28 have B DIF. An analysis of variance (ANOVA) of the past 400 MH D-DIF values shows that for the male-female DIF analysis, the main effects of D1 and D2 classifications are significant at the 5% level and a backward stepwise algorithm chooses a model with the main effects of D1 and D2 classifications. For the White-Black DIF analysis, only the main effect of the D2 classification is significant at the 5% level, and a backward stepwise algorithm chooses a model with the main effect of D2 classification only.

Table 2 shows the mean and standard deviation (SD) of the MH D-DIF values for the 400 past items for different item types. These values will act as the prior means and SDs ($\widehat{\mu_T}$ and $\sqrt{\widehat{\tau_T^2}}$) for the FB analysis.

5.2 Results From DIF Analysis

Both of the two most recent PPSTR administrations have items with all four combinations of D1 and D2 classifications. Table 3 summarizes the performance of the different statistics for male-female and White-Black DIF for the two most recent PPSTR data sets.

Four versions of the FB estimate were considered:

Table 2.
Mean and SD of the MH D-DIF Statistics for Different Item Types for the Past 10 Administrations of the PPSTR Assessment

D1 classification	D2 classification	Number of items	Male-female DIF		White-Black DIF	
			Mean	SD	Mean	SD
1	1	177	0.07	0.45	0.10	0.55
1	2	54	-0.07	0.44	-0.22	0.49
2	1	123	0.00	0.47	0.13	0.49
2	2	46	-0.23	0.47	-0.28	0.57

1. *FB (D1 & D2)* denotes a FB estimate using four item types (i.e., four $\widehat{\mu_T}$s and four $\widehat{\tau_T^2}$s), one for each combination of D1 and D2 classification.

2. *FB (D1)* denotes a FB estimate using item types based on D1 classification only.

3. *FB (D2)* denotes a FB estimate using item types based on D2 classification only.

4. The fourth version, *FB (1),* denotes an FB estimate using one item type that contains all items; this estimate is very much what the quote in Section 3. from Zwick et al. (1999, p. 7) referred to. For this version, $\widehat{\mu_T}$ and $\widehat{\tau_T^2}$ are the mean and variance of the MH D-DIF statistics of the 400 past items.

The results for the first three above were virtually indistinguishable—so this paper reports results for only one of these, depending on an ANOVA of the MH D-DIF values; for example, if an ANOVA suggests that effects of both D1 and D2 classifications are significant, results for FB (D1 & D2) are reported. The leftmost column in Table 3 shows the data set (i.e., the administration [1 or 2] the data are from, focal group size, reference group size, the number of C DIF items, and the number of B DIF items). There are few items with B or C DIF categories in the two current data sets. The remaining columns in the table show, for each condition and each method, the (a) root average squared bias computed over all items, (b) average RMSE computed over all items, (c)-(e) average percent correct DIF categorization (with respect to keep/flag) for keep items, flag items, and overall, and (f)-(g) average loss for keep items, flag items, and overall. Section 4. describes how these measures are computed. Results for the B rule, C rule, and EB method are also reported, rather than those for the two versions of the FB method. Note that given the choice of 1 as the indifference point in the loss function, as in Zwick et al. (2000), the results for the EB

Table 3.
The Outcome for Male-Female and White-Black DIF for PPSTR

Simul. cond.	Method	Root av. sq. bias	Av. RMSE	Av. % corr. keep	Av. % corr. flag	Av.% corr. all	Av. loss keep	Av loss flag	Av. loss all
MF	C rule	0.21	1.28	98.43	8.43	91.68	0.18	1.58	0.29
Data 1	B rule	0.21	1.28	88.19	29.70	83.80	0.26	1.39	0.35
186	EB	0.41	0.49	91.67	16.40	86.03	0.24	1.53	0.33
50	FB (D1 & D2)	0.45	0.39	100.0	0.00	92.50	0.17	1.66	0.28
1, 2	FB (1)	0.40	0.40	100.0	0.00	92.50	0.17	1.66	0.28
MF	C rule	0.02	0.44	99.75	18.50	93.65	0.17	1.42	0.27
Data 1	B rule	0.02	0.44	90.29	70.57	88.81	0.23	1.08	0.29
1120	EB	0.24	0.32	99.00	25.33	93.48	0.18	1.36	0.27
300	FB (D1 & D2)	0.26	0.31	99.79	16.10	93.52	0.17	1.43	0.27
1,2	FB (1)	0.27	0.31	99.91	15.13	93.55	0.17	1.44	0.27
MF	C rule	0.09	1.29	98.46	-	98.46	0.19	-	0.19
Data 2	B rule	0.09	1.29	89.07	-	89.07	0.26	-	0.26
177	EB	0.35	0.42	94.92	-	94.92	0.22	-	0.22
50	FB (D1 & D2)	0.36	0.35	100.00	-	100.00	0.18	-	0.18
0, 0	FB (1)	0.37	0.35	100.00	-	100.00	0.18	-	0.18
WB	C rule	0.08	1.08	98.56	6.92	84.81	0.22	1.40	0.40
Data 1	B rule	0.08	1.08	87.48	30.28	78.90	0.30	1.29	0.45
50	EB	0.45	0.54	92.36	17.40	81.12	0.27	1.35	0.43
263	FB (D2)	0.47	0.45	99.99	0.22	85.02	0.21	1.43	0.40
0,6	FB (1)	0.51	0.48	99.99	0.00	84.99	0.21	1.43	0.40
WB	C rule	0.03	0.36	99.66	11.65	86.46	0.22	1.37	0.39
Data 1	B rule	0.03	0.36	91.60	71.60	88.60	0.25	1.11	0.38
300	EB	0.21	0.32	97.84	35.65	88.51	0.22	1.26	0.38
1583	FB (D2)	0.24	0.31	99.17	27.23	88.38	0.22	1.30	0.38
0,6	FB (1)	0.24	0.32	99.22	24.02	87.94	0.22	1.31	0.38
WB	C rule	0.07	1.08	98.79	6.72	87.28	0.19	1.60	0.37
Data 2	B rule	0.07	1.08	88.29	28.30	80.79	0.27	1.47	0.42
50	EB	0.46	0.50	94.45	13.06	84.27	0.22	1.56	0.39
258	FB (D2)	0.48	0.44	99.98	0.18	87.51	0.18	1.65	0.36
1, 4	FB (1)	0.49	0.45	100.00	0.02	87.50	0.18	1.65	0.36

Note. The Simul. cond. column gives the type of DIF analysis done (i.e., MF for male-female or WB for White-Black), data set (1 or 2), number of examinees in the focal and reference groups, number of true C DIF items, and number of true B DIF items according to the ETS DIF rule.

and FB method will be compared to those from the B rule (and not to those from the C rule) in the remainder of this paper. Thus *MH estimate* refers to the B rule henceforth. However, results for the C rule, the operational method, are also reported.

Table 3 shows that the FB method performs the best for all the cases with respect to RMSE, proportion correct decisions, and average loss. The results for the FB(1) version are very close to those for FB (D1 and D2), FB(D1), or FB(D2), which shows that whether the FB analysis considers one type of items or more does not affect the results. The FB analysis results in substantial gains over the other methods for small sample sizes. For example, for Data Set 1, the RMSE for the FB method is 0.39, while the lowest among the others is 0.49. For the same data set, the average overall loss for the FB method is 0.28, while the lowest among the B rule and the EB method is 0.33. Like the EB method, the FB method provides biased estimates as well, but it is more than compensated by its improved RMSE.

The appendix shows results of the same type of analyses as above on PPST mathematics and PPST writing. These results are very similar to those for the PPSTR (i.e., the FB estimate performs slightly better than the EB estimate, which performs better than the MH D-DIF estimate). However, the gain for the FB method over the EB method is not substantial for a number of cases.

We also performed one more set of simulations for PPSTR data; these simulations are mostly similar to the above mentioned ones, but they differ in that we now divided the data set for each of the two most recent test forms into two, depending on the time an examinee took the test (it so happens that each of these test forms were given on two different dates). We drew a random small subsample from the data with all the examinees who took the test at the earlier test date, computed DIF estimates and classifications, and compared them to the corresponding values computed from the full data set with all the examinees who took the test at the later test date. Table 4 provides bias, RMSE, percent correct decision, average loss, and so on, for the estimates.

The results from Table 4 are often similar to the earlier results (e.g., the EB and FB methods have less RMSE than the MH method). The major difference is that the MH statistic has comparatively more bias than the earlier DIF analyses; as a result, the bias of the EB and FB estimates are is much worse than that for MH estimates. Besides, the EB and FB methods appear virtually indistinguishable with respect to the values of the measures.

We investigated further the results reported in Table 3 to understanding better the nature of

Table 4.

The Outcome for Male-Female and White-Black DIF When DIF Estimates From an Earlier Time-Point Are Used to Predict DIF at a Later Time-Point for PPSTR

Simul. cond.	Method	Root av. sq. bias	Av. RMSE	Av. % corr. keep	Av. % corr. flag	Av.% corr. all	Av. loss keep	Av loss flag	Av. loss all
MF	C rule	0.46	1.32	98.28	5.98	89.05	0.18	1.50	0.31
Data 1	B rule	0.46	1.32	87.96	23.50	81.51	0.27	1.37	0.38
189	EB	0.50	0.43	99.75	0.00	89.77	0.17	1.56	0.31
50	FB (D1 & D2)	0.49	0.43	100.00	0.00	90.00	0.17	1.56	0.31
1,3	FB (1)	0.50	0.43	100.00	0.00	90.00	0.17	1.56	0.31
MF	C rule	0.40	1.32	98.61	-	98.61	0.21	-	0.21
Data 2	B rule	0.40	1.32	89.04	-	89.04	0.29	-	0.29
180	EB	0.41	0.38	99.84	-	99.84	0.21	-	0.21
50	FB (D2)	0.40	0.37	100.00	-	100.00	0.20	-	0.20
0,0	FB (1)	0.41	0.38	100.00	-	100.00	0.20	-	0.20
WB	C rule	0.38	1.10	98.02	6.63	91.17	0.24	1.38	0.33
Data 1	B rule	0.38	1.10	85.18	28.13	80.90	0.33	1.30	0.40
50	EB	0.46	0.45	99.82	0.07	92.34	0.23	1.41	0.32
266	FB (D2)	0.43	0.42	99.94	0.03	92.45	0.23	1.41	0.32
0,3	FB (1)	0.46	0.45	99.99	0.07	92.50	0.23	1.41	0.32
WB	C rule	0.34	1.20	98.82	4.68	87.06	0.17	1.54	0.34
Data 2	B rule	0.34	1.20	89.45	22.14	81.03	0.24	1.45	0.39
50	EB	0.50	0.46	99.86	0.00	87.38	0.16	1.57	0.33
233	FB (D2)	0.49	0.44	99.98	0.06	87.49	0.16	1.57	0.33
1,4	FB (1)	0.50	0.45	100.00	0.00	87.50	0.16	1.57	0.33

Note. The Simul. cond. column gives the type of DIF analysis done (i.e., MF for male-female or WB for White-Black), data set (1 or 2), number of examinees in the focal and reference groups, number of true C DIF items, and number of true B DIF items according to the ETS DIF rule.

the FB approach. Figure 3 shows, for male-female DIF analysis with a focal group size of 50 and a reference group size of 186 in PPSTR Data Set 1, plots of the true values of the MH statistic (which are the values of the MH statistics from the full data set) versus the bias (top panel) and the RMSE (bottom panel) for the MH estimate, EB estimate, and FB estimate.

Figure 3 also shows that the EB and FB estimates are more biased than the MH estimates in general. This is because of shrinkage to the mean (see, e.g., Section 2.2). The figure also shows that, overall, the EB estimates have much less RMSE than the MH estimates. However, the FB

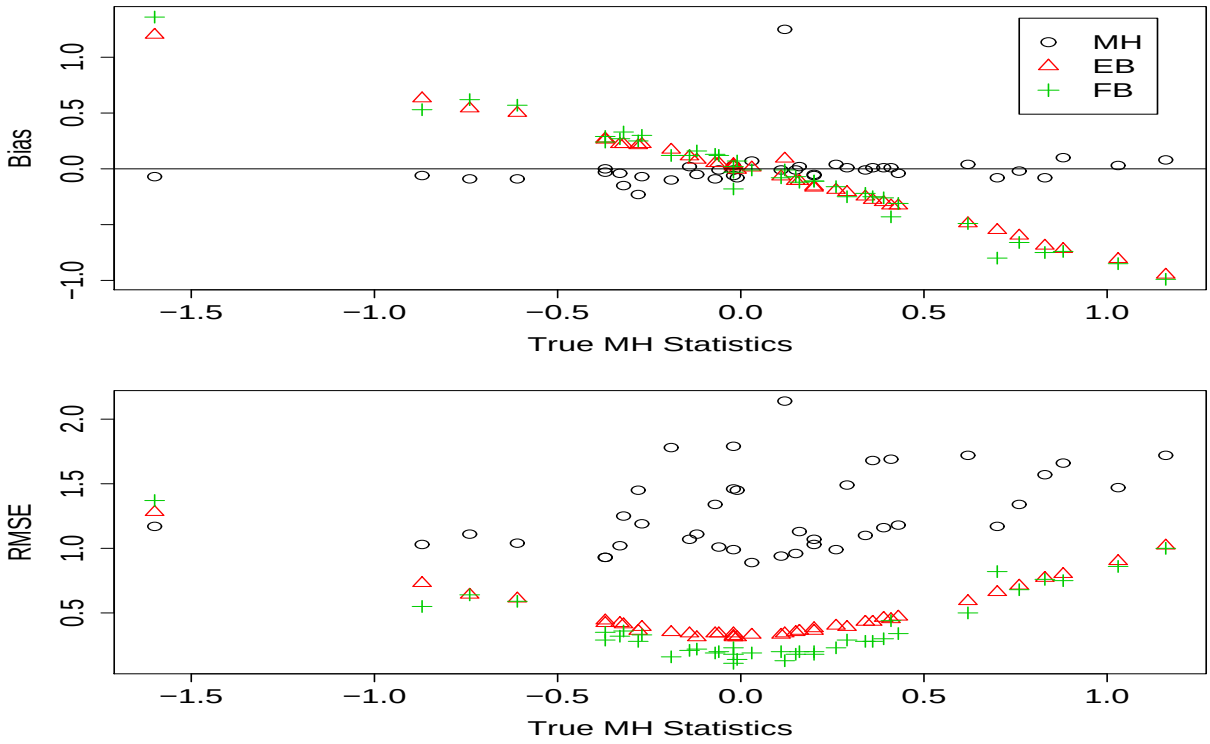

***Figure* 3. Bias and MSE for the three estimates for PPSTR for male-female DIF.**

estimates have even less RMSE than the EB estimates for all but two or three items.

For PPSTR Data Set 1, Figure 4 compares the percentage of times an item got flagged (top panel) and the average loss functions (bottom panel) given by Equation 11 for the three estimates for male-female DIF.

The top plot in Figure 4 shows that the FB method never flags an item for this case; that makes the FB method the best method for the items that are actual keep items, but that also means that the FB method will lead to an error for the actually flag items. The consequences are observed in the bottom plot of Figure 4, which shows that the average loss for the FB estimate is lower than that for the MH and EB estimates in the middle, while the average loss for the FB estimate is higher than that for the other two estimates for high or low values of the true DIF parameter. The opposite is true for the MH estimate. The average loss of the EB estimate lies in between the other two.

The EB estimate is slightly more conservative than the MH estimate regarding DIF classification as also observed in Zwick et al. (2000). However, the FB estimate seems to be more

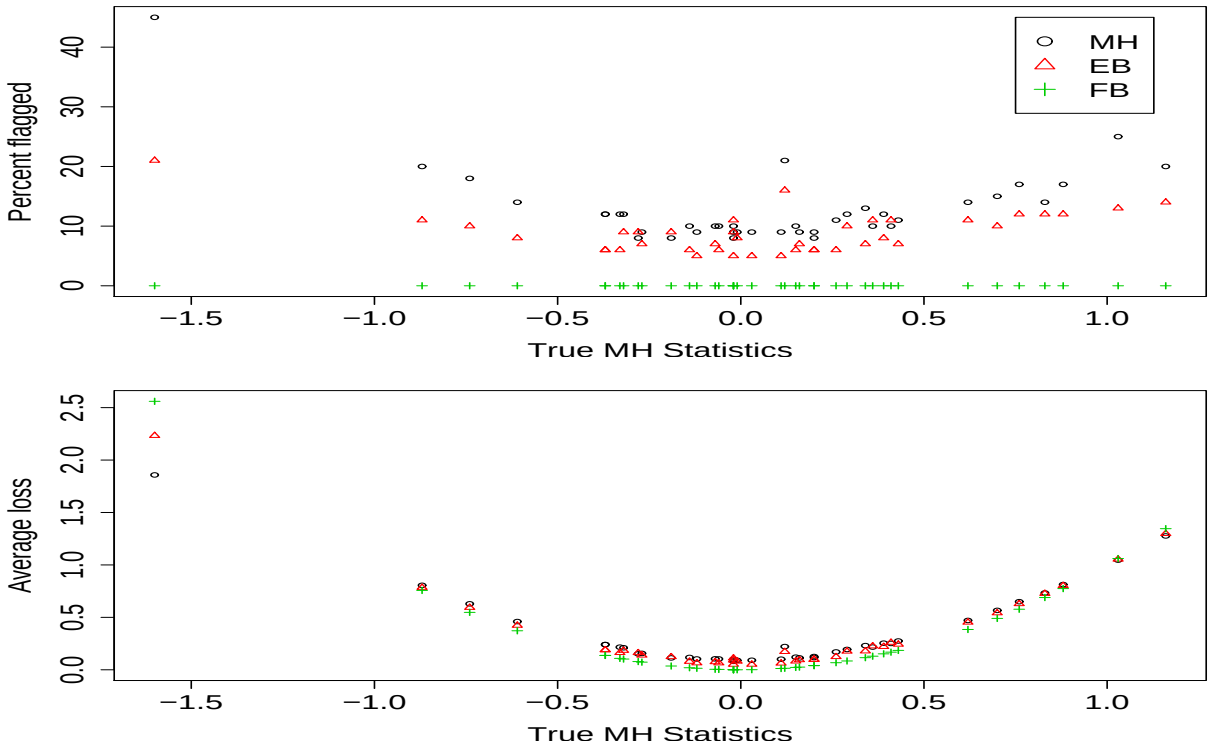

***Figure* 4. Average loss for the three estimates for PPSTR for male-female DIF.**

conservative than the MH and EB estimates (it rarely flags an item, as can be seen from the top panel of Figure 4). To examine the classification issue further, Figure 5 compares, for PPSTR Data Set 1 and two items,[2] plots of the difference of the male-female MH/EB/FB estimates and the true MH D-DIF value for a focal group size of 50 and a reference group size of 186.

Figure 6 compares the weights assigned to the MH D-DIF estimate by the EB estimate and the FB estimate for the two items in the 1,000 replications; the expression for the weight for the EB estimate is given by Equation 7, while that for the FB estimate is given by Equation 15. The figure shows that for both items, the weights for the FB estimate are all just below 0.2, while those for the EB estimate are between 0 to 0.7. The reason is that the EB estimates of τ^2 (with a median of 0.34) are mostly larger than the FB estimates of τ_T^2 (around 0.18 for both items) for the items.

For small sample sizes, the S_i^2 (or S_{iT}^2)s are large—with 2.5th percentile, median, and 97.5th percentile of the distribution of S_i^2 (or S_{iT}^2) in the 1,000 subsamples for the two items as $\{0.79, 1.00, 1.41\}$ and $\{0.67, 0.83, 1.13\}$, respectively, which leads to values of weight being

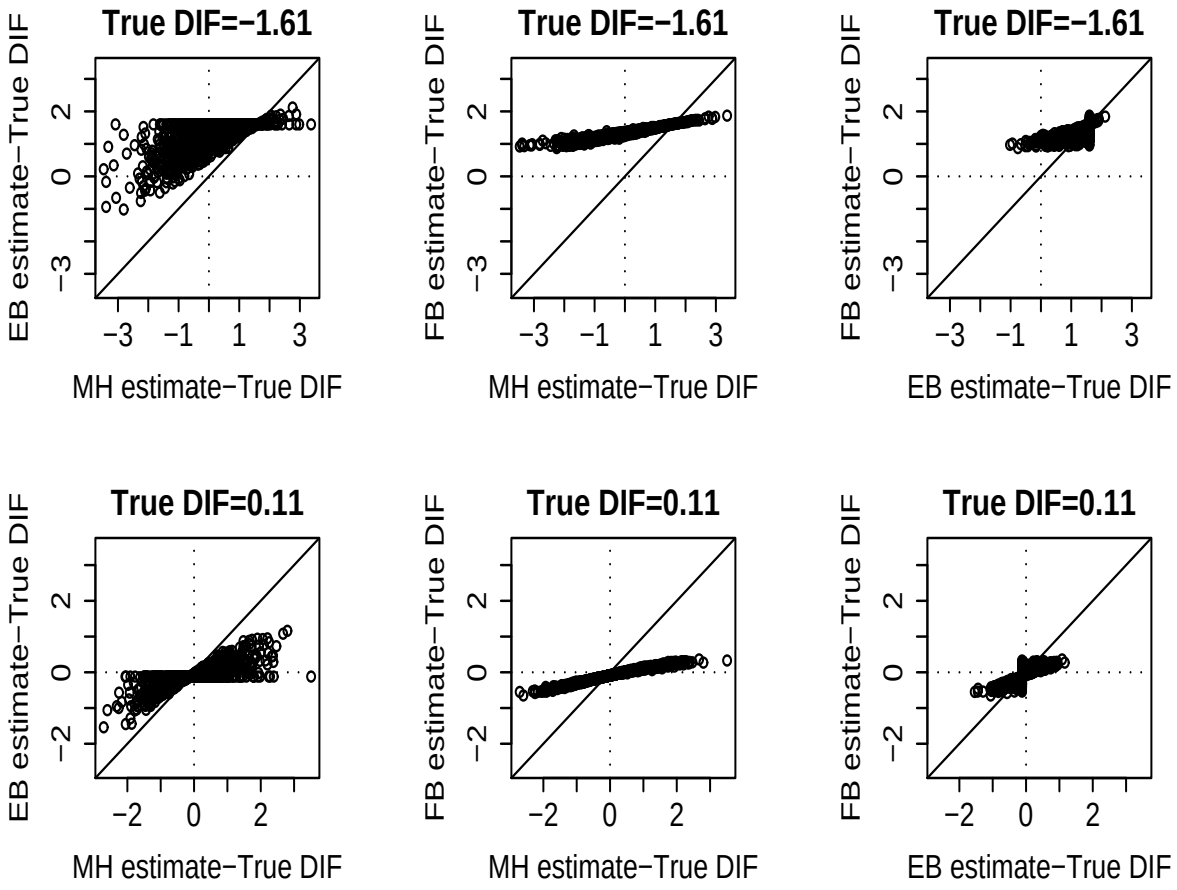

***Figure* 5. Plot of estimates minus true DIF for different methods for two items (each row showing plots for an item) for PPSTR for male-female DIF.**

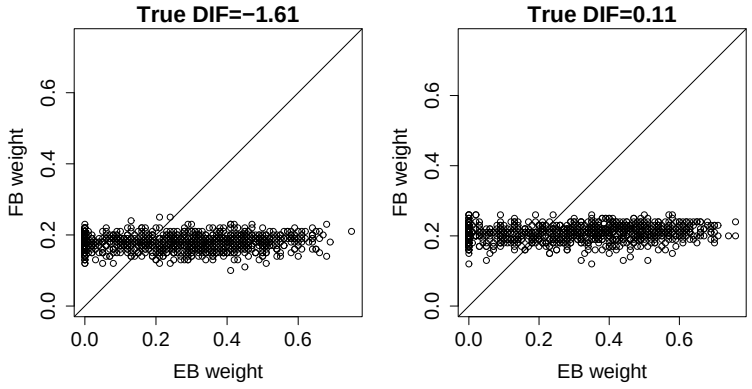

***Figure* 6. Plot of EB weights versus the FB weights for two items for PPSTR for male-female DIF.**

considerably less than 1 for EB, and even lesser for FB. Hence Figure 6, together with Equations 6 and 14 (which imply that the amount of shrinkage to the mean increases as weight decreases), suggests that the EB estimates are shrunk to the prior mean (which is close to 0) to a certain extent, but the FB estimates are shrunk even more. For the second item (with true DIF category A and true DIF value of 0.11, which is close to 0), the shrinkage leads to lower RMSE for the FB estimate than for the EB and MH estimates. For the first item, which has true DIF category C, however, the true DIF (-1.61) is far from 0, and hence the shrinkage causes the Bayesian estimates (both EB and FB) to be consistently far from the true value and to have higher RMSE than the MH estimate. The correlation coefficients between the MH and EB estimates are 0.64 and 0.80 for the two items; those between the MH and FB estimates are 0.97 and 0.99 (because, from Equation 14 and Figure 6, it is clear that the FB estimate is always close to $0.2 * \text{MH estimate} + 0.8 * \widehat{\mu_T}$, resulting in a high correlation between these two estimates) and those between the EB and FB estimates are 0.63 and 0.80.

One more interesting feature of the results shown in Table 3 is that the performance of the C rule with respect to percent correct classifications and average loss is often very close to that of the EB estimate and more so to that of the FB estimate (the average RMSE of the C rule, however, is always much larger than that for the EB and FB estimates). Section 2.2 discussed that the EB and FB estimates are comparable directly to the B rule because of the choice of an indifference point of 1. Because of its conservative nature, however, the FB estimate (and, to a certain extent, the EB estimate) behaves more like a C rule than a B rule with respect to DIF classification.

We also performed the above analyses with the indifference point of 1.5; in this case, the EB and FB estimates become comparable to the C rule, rather than to the B rule. Note that the values for the RMSE does not change if the indifference point is changed, so the FB estimate is slightly better than the EB estimate, which is slightly better than the C rule with respect to the average loss. The values regarding the classification and average loss are different from those obtained with indifference point of 1, but still the EB and FB estimates perform slightly better than the C rule, and the EB estimate performs almost as well as the FB estimate with respect to those values.

6 Discussion and Conclusions

This paper suggests an FB approach to DIF analysis using data from past operational forms of the test concerned to form a prior distribution. Thus, this paper is the first to make an attempt to use information from past DIF analyses to improve on current DIF analysis. The suggested approach performs better than the existing ones for small samples, but the gain is not too substantial overall. In particular, the detailed information on different item types was not found to be very useful. For almost all cases, an FB analysis with only one item type performs as well as a FB analysis with more item types. This is because (a) the variation in the DIF estimates in the past data is too low (there are very few B or C DIF items in the past), causing too much shrinkage toward the overall mean of the small-sample estimates, resulting in very few flag decisions, and, (b) the most recent data sets have few items with substantial DIF (so that any approach that never flags an item will not perform too poorly). One might wonder if the superior performance of the FB estimate compared to the other estimates can be attributed to shrinkage to the mean only. In other words, will an estimate that results in more shrinkage to the mean always perform better? The answer is no. For example, if one considers extreme shrinkage in the form of a DIF rule that always estimates the DIF parameters of all items by 0, its RMSE for male-female DIF for Data Set 1 for a focal group size of 50 is 0.53; the EB estimate (with RMSE of 0.49) beats this trivial rule barely, while the FB estimate (with RMSE of 0.39) beats the rule by a considerable margin; the average loss for the rule is the same as that for the FB estimate.

Figure 3 shows that the EB and FB estimates have higher RMSE than the MH estimate for the one extreme true MH value. The poor performance of Bayesian estimates compared to frequentist estimates for extreme observations have been noted in, for example, Carlin and Louis (1996, p. 88). The fact that the most recent data sets analyzed in these studies have very few items with extreme DIF may have mostly contributed to the EB and FB approaches appearing better overall than the MH approach (though one could argue that because few of the past items have B or C DIF, it is expected that the recent data will have the same feature). In the future, we would like to examine the performance of the FB estimate for an assessment at the other end of the spectrum; that is, for which the proportion of items with B or C DIF is much more than what is observed here.

From a practical perspective, the results of this study provide some support for the current practice of performing no DIF analysis under small sample sizes for tests that have a prior history

of finding only a few items flagged for DIF. Performing no DIF analyses at all is consistent with using a very strong prior in which all items are presumed to be A DIF items. When the prior history reveals a preponderance of A DIF items, this presumption of all A DIF items is a sound wager from a statistical perspective. It remains to be seen how effective the full Bayesian method is with tests where DIF is more prevalent.

This work leaves a number of issues for further research. Though this paper considered the Mantel-Haenszel statistic only, the method suggested can also be applied to other DIF-detection methods (e.g., SIBTEST, standardization, or logistic regression). One could perform a more thorough Bayesian analysis using hyper-prior distributions on μ_T and τ_T^2 or allowing σ_{iT}^2s to be random parameters—the computational burden will increase for these analyses, though. It is also possible to include the past information in some other way than the one used in this paper, such as using a power prior distribution (Ibrahim & Chen, 2000); a power prior distribution is based on constructing the likelihood function from the past data and raising it to a scalar power between 0 and 1 to account for the difference between the current and past data. One could also explore indifference points other than 1 with the loss function of Zwick et al. (2000) and explore other types of loss functions.

References

Braun, H. I. (1989). Empirical Bayes methods: A tool for exploratory analysis. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 19-55). San Diego, CA: Academic Press Inc.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement, 34*, 123-139.

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

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis.* London: Chapman & Hall.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

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: Lawrence Erlbaum Associates.

Dorans, N. J., Schmitt, A., & Curley, W. E. (1988). *Differential speededness: Some items have DIF because of where they are, not because of what they are.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75*, 165-190.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* New York: Chapman & Hall.

Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. *Journal of Educational Measurement, 40*, 281-306.

Holland, P. W. (1985). On the study of differential item performance without IRT. *Proceedings of the 27th annual conference of the Military Testing Association* (Vol. I; pp. 282-287). San Diego, CA: Navy Personnel Research and Development Center.

Holland, P. W. (1987a). *Expansion and comments on Marco's rational approach to flagging items for DIF* (ETS internal memorandum). Princeton, NJ: ETS.

Holland, P. W. (1987b). *More on rational approach item flagging* (ETS internal memorandum). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (ETS RR-85-43). Princeton, NJ: ETS.

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

Ibrahim, J. G., & Chen, M. H. (2000). Power prior distributions for regression models. *Statistical Sciences, 15,* 46-60.

James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability: Vol. I. Contributions to the theory of statistics* (pp. 361-380). Berkeley, CA: University of California Press.

Johnson, W. O., & Gastwirth, J. (1991). Bayesian inference for medical screening tests: Approximations useful for the analysis of AIDS data. *Journal of the Royal Statistical Society, Series B, 53*, 427-439.

Lawrence, I. M., & Curley, W. E. (1989, April). *Differential item functioning for males and females on SAT–Verbal reading subscore items: Follow-up study.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Lawrence, I. M., Curley, W. E., & McHale, F. J. (1988). *Differential item functioning for males and females on SAT–Verbal reading subscore items* (ETS RR-88-04). Princeton, NJ: ETS.

Lewis, C., & Thayer, D. T. (2000). *An investigation of the effectiveness of Mantel-Haenszel procedures to detect DIF using pretest data from an adaptive licensure test.* Unpublished manuscript.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*(2), 157-162.

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.

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004). Differential item functioning methods based on small samples: An illustration based on an attitude survey. *Journal of Educational Measurement, 41*, 331-344.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Phillips, A., & Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics, 43*, 425-431.

Robbins, H. I. (1955). An empirical Bayes approach to statistics. *Proceedings of the third Berkeley symposium on mathematical statistics and probability: Vol 1. Contributions to the theory of statistics* (pp. 157-164). Berkeley, CA: University of California Press.

Schmitt, A. P. (1985). *Assessing unexpected differential item performance of Hispanic candidates on SAT form 3FSA08 and TSWE form E47* (ETS SR-85-169). Princeton, NJ: ETS.

Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic aptitude test. *Journal of Educational Measurement, 25*, 1-13.

Schmitt, A. P., & Bleistein, C. A. (1987). *Factors affecting differential item functioning of Black examinees on Scholastic Aptitude Test analogy items* (ETS RR-87-23). Princeton, NJ: ETS.

Schmitt, A. P., Curley, W. E., Bleistein, C. A., & Dorans, N. J. (1988). *Experimental evaluation of language and interest factors related to differential item functioning for Hispanic examinees on the SAT-Verbal.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27*, 67-81.

Schmitt, A. P., Dorans, N. J., & Holland, P. W. (1993). Evaluating hypothesis about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sinharay, S. (2005). *Bayesian statistics in educational measurement.* Paper presented at the international workshop/conference on Bayesian statistical analysis, Varanasi, India.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors. *Applied*

Psychological Measurement, 27, 27-31.

Tsutakawa, R. K. (1992). Prior distribution for item response curves. *British Journal of Mathematical and Statistical Psychology, 45*, 201-374.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*, 339-368.

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*(1), 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.

Notes

[1] Zwick et al., 2000, p. 229, commented that Parshall and Miller (1995) found the performance of the usual MH procedure to be similar to that of an exact MH test under various conditions, including focal group samples as small as 25.

[2] The two items both have D1 classification of 2 and D2 classification of 1. One item (the leftmost item in Figure 3) has a true MH D-DIF value of -1.61, and the other item has a true MH D-DIF value of 0.11.

Appendix

This appendix describes the results from same type of analyses as above for PPST mathematics (PPSTM) and PPST writing (PPSTW) tests.

Analyses of PPSTM Data

Analysis of Past Data

For the PPST mathematics (PPSTM) assessment, test developers classify each item into one of five D1 classifications with respect to content area. Each item is also classified into one of two D2 classifications.

Figures A1 and A2, respectively, show the distribution of the male-female and White-Black MH D-DIF statistics for 400 past items from 10 past data forms of the PPSTM assessment according to the different classifications. For male-female DIF, 5 of the past 400 items have C DIF and 25 have B DIF. For White-Black DIF, 3 of the past 400 items have C DIF and 32 have B DIF.

An ANOVA of the MH D-DIF values shows that for the male-female DIF analysis, the main effects for the two classifications and their interaction effect are significant at the 5% level, while for the White-Black DIF analysis, only the main effect for the D1 classification is significant at the 5% level.

For male-female DIF, a backward stepwise algorithm results in a model with the main effects for the two classifications and their interaction effect; for White-Black DIF, a backward stepwise algorithm results in a model with main effect of the D1 classification only.

Results From DIF Analysis

Both recent administrations have items with all combinations of D1 and D2 classifications. Table A1 summarizes the performance of the different statistics for male-female and White-Black DIF. As with reading, there are only a few items with B or C DIF.

Table A2 shows the results of the same type of analyses performed for Table A1.

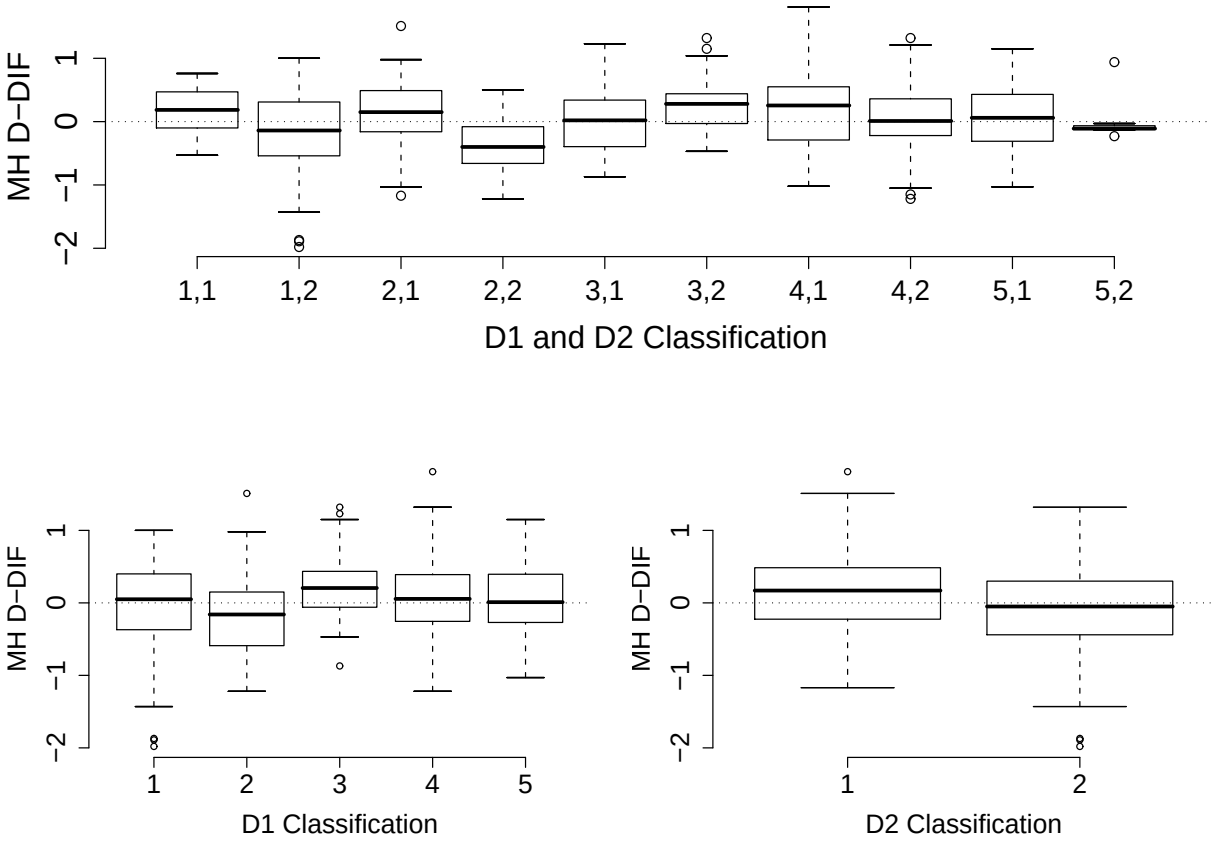

Figure A1. ***Distribution of the MH D-DIF statistics for PPSTM for male-female DIF.***

Note. The dotted line corresponds to an MH D-DIF value of 0.

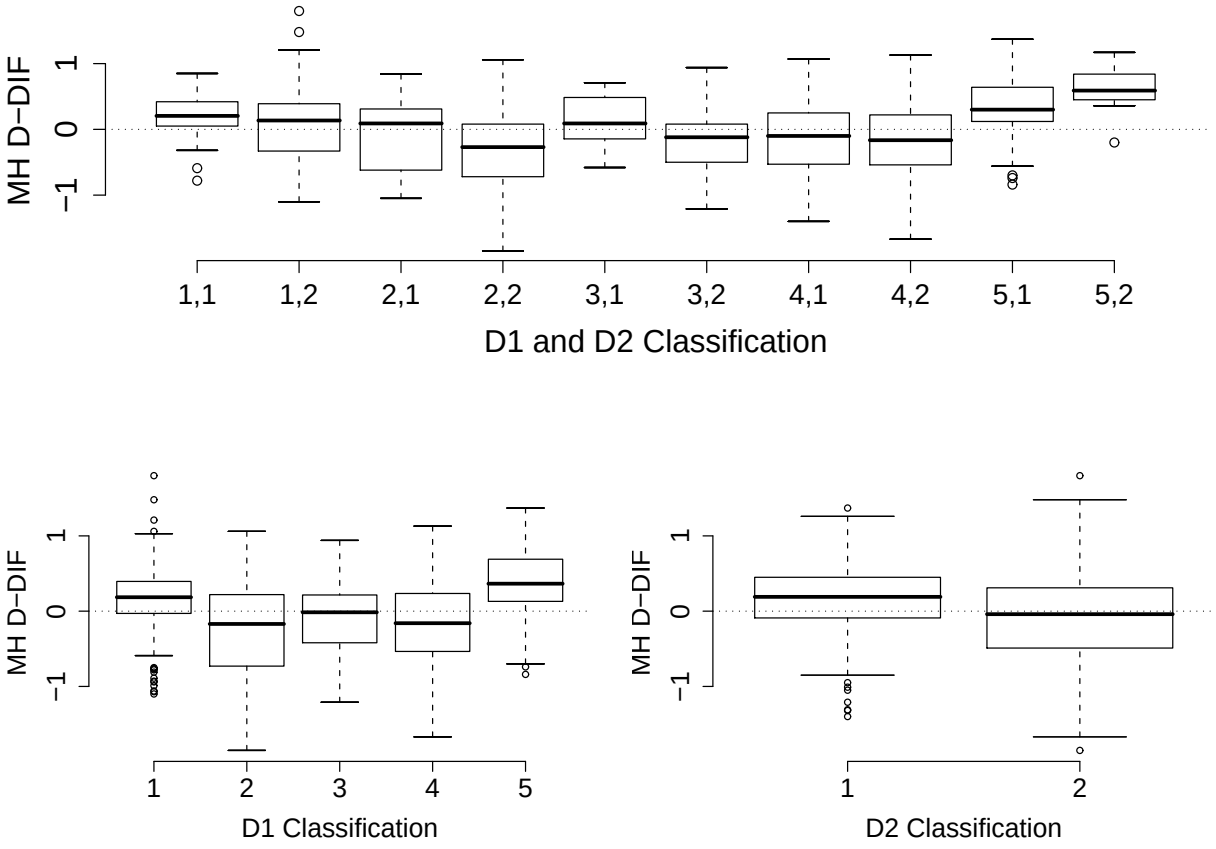

Figure A2. ***Box plots for the MH D-DIF statistics for PPSTM for White-Black DIF.***

Note. The dotted line corresponds to an MH D-DIF value of 0.

Table A1

The Outcome for Male-Female and White-Black DIF for PPSTR

Simul. cond.	Method	Root av. sq. bias	Av. RMSE	Av. % corr. keep	Av. % corr. flag	Av.% corr. all	Av. loss keep	Av loss flag	Av. loss all
MF	C rule	0.07	1.16	98.34	8.37	91.59	0.23	1.68	0.34
Data 1	B rule	0.07	1.16	86.94	34.00	82.97	0.31	1.39	0.39
186	EB	0.43	0.50	92.42	16.50	86.73	0.27	1.62	0.37
50	FB (D1 & D2)	0.48	0.45	99.90	0.00	92.41	0.22	1.81	0.34
1, 2	FB (1)	0.45	0.44	99.99	0.47	92.53	0.22	1.80	0.34
MF	C rule	0.02	0.40	99.64	27.20	94.20	0.22	1.34	0.30
Data 1	B rule	0.02	0.40	89.29	72.23	88.01	0.27	1.06	0.33
1119	EB	0.19	0.32	97.67	39.80	93.33	0.23	1.23	0.30
300	FB (D1 & D2)	0.25	0.32	98.83	18.10	92.77	0.22	1.46	0.32
1, 2	FB (1)	0.20	0.31	98.58	37.27	93.98	0.22	1.21	0.30
MF	C rule	0.21	1.24	98.43	10.02	87.38	0.25	1.92	0.46
Data 2	B rule	0.21	1.24	88.55	37.14	82.13	0.32	1.54	0.47
185	EB	0.53	0.57	90.33	20.78	81.64	0.31	1.85	0.50
50	FB (D1 & D2)	0.60	0.50	99.94	0.32	87.49	0.24	2.19	0.48
1, 4	FB (1)	0.58	0.51	100.00	0.02	87.50	0.24	2.19	0.48
WB	C rule	0.07	1.06	98.70	8.82	87.47	0.19	1.64	0.37
Data 1	B rule	0.07	1.06	87.85	35.30	81.28	0.27	1.37	0.41
50	EB	0.44	0.51	94.40	15.74	84.57	0.22	1.60	0.39
240	FB (D1)	0.47	0.46	99.86	6.34	88.17	0.18	1.62	0.36
1, 4	FB (1)	0.47	0.46	99.96	1.28	87.63	0.18	1.75	0.38
WB	C rule	0.04	0.37	99.89	20.38	89.95	0.18	1.30	0.32
Data 1	B rule	0.04	0.37	93.04	66.12	89.68	0.22	1.09	0.33
300	EB	0.20	0.31	98.79	36.80	91.04	0.19	1.21	0.31
1441	FB (D1)	0.21	0.31	99.48	36.30	91.58	0.18	1.17	0.31
1, 4	FB (1)	0.20	0.31	99.27	35.74	91.33	0.18	1.19	0.31
WB	C rule	0.11	1.06	98.85	6.73	89.64	0.18	1.39	0.30
Data 2	B rule	0.11	1.06	87.77	29.23	81.92	0.26	1.29	0.37
50	EB	0.42	0.47	95.14	11.88	86.82	0.21	1.37	0.33
328	FB (D1)	0.38	0.40	99.85	3.65	90.23	0.17	1.41	0.30
0, 4	FB (1)	0.42	0.43	99.91	0.93	90.01	0.17	1.42	0.30

Note. The Simul. cond. column gives the type of DIF analysis done (i.e., MF for male-female or WB for White-Black), data set (1 or 2), number of examinees in the focal and reference groups, number of true C DIF items, and number of true B DIF items according to the ETS DIF rule.

Table A2

The Outcome for Male-Female and White-Black DIF When DIF Estimates From an Earlier Time-Point Are Used to Predict DIF at a Later Time-Point for PPSTR

Simul. cond.	Method	Root av. sq. bias	Av. RMSE	Av. % corr. keep	Av. % corr. flag	Av.% corr. all	Av. loss keep	Av loss flag	Av. loss all
MF	C rule	0.34	1.17	98.43	9.80	91.78	0.24	1.68	0.35
Data 1	B rule	0.34	1.17	87.77	37.13	83.97	0.32	1.37	0.40
187	EB	0.48	0.47	99.65	0.27	92.20	0.24	1.82	0.35
50	FB (D1 & D2)	0.50	0.46	99.93	0.00	92.44	0.23	1.82	0.35
1,2	FB (1)	0.48	0.46	99.99	0.27	92.51	0.23	1.82	0.35
MF	C rule	0.41	1.23	98.64	7.80	85.02	0.23	1.90	0.48
Data 2	0.41	1.23	88.19	25.62	78.80	0.31	1.60	0.50	
182	EB	0.61	0.53	99.64	0.00	84.69	0.23	2.12	0.51
50	FB (D1 & D2)	0.64	0.52	99.93	0.00	84.94	0.23	2.12	0.51
1,5	FB (1)	0.61	0.53	100.00	0.00	85.00	0.23	2.12	0.51
WB	C rule	0.34	1.04	98.52	18.45	94.52	0.22	1.70	0.29
Data 1	B rule	0.34	1.04	86.90	50.75	85.09	0.30	1.33	0.35
50	EB	0.41	0.43	99.77	3.50	94.96	0.21	1.92	0.29
247	FB (D1 & D2)	0.41	0.43	99.84	19.95	95.85	0.21	1.64	0.28
1,1	FB (1)	0.41	0.42	99.96	3.50	95.14	0.21	1.92	0.29
WB	C rule	0.56	1.06	99.12	1.97	91.83	0.20	1.66	0.31
Data 2	B rule	0.56	1.06	87.92	15.50	82.49	0.29	1.57	0.38
50	EB	0.48	0.48	99.85	0.23	92.38	0.20	1.68	0.31
378	FB (D1 & D2)	0.45	0.44	99.71	1.83	92.37	0.20	1.66	0.31
1,2	FB (1)	0.49	0.48	99.90	0.23	92.43	0.20	1.68	0.31

Note. The Simul. cond. column gives the type of DIF analysis done (i.e., MF for male-female or WB for White-Black), data set (1 or 2), number of examinees in the focal and reference groups, number of true C DIF items, and number of true B DIF items according to the ETS DIF rule.

Analyses of PPSTW Data

Analysis of Past Data

For the PPST writing (PPSTW) assessment, test developers classify each item into 1 of 4 D1 classifications. Each item is also classified into 1 of 17 D2 classifications.

Figures A3 and A4, respectively, show the distribution of the male-female and White-Black MH D-DIF statistics for 442 items from 10 previous forms of the PPSTW assessment according to the different classifications. For male-female DIF, none of the 442 items have C DIF and 9 have B DIF. For White-Black DIF, 8 of the items have C DIF and 25 have B DIF.

An ANOVA of the MH D-DIF values shows that for both male-female and White-Black DIF analyses, the main effects for the two classifications are significant at the 5% level. For male-female DIF, a backward stepwise algorithm results in a model with main effect of the D1 classification only; for White-Black DIF, a backward stepwise algorithm results in a model with main effect of the D2 classification only.

Results From DIF Analysis

Both recent administrations have items with all combinations of D1 classifications and D2 classifications except for $\{4, 5\}$. Table A3 summarizes the performance of the different statistics for male-female and White-Black DIF. There are very few items with B or C DIF categories.

Both recent administrations have items with all combinations of D1 and D2 classifications except for $\{4, 5\}$. Table A3 summarizes the performance of the different statistics for male-female and White-Black DIF. There are very few items with B or C DIF categories.

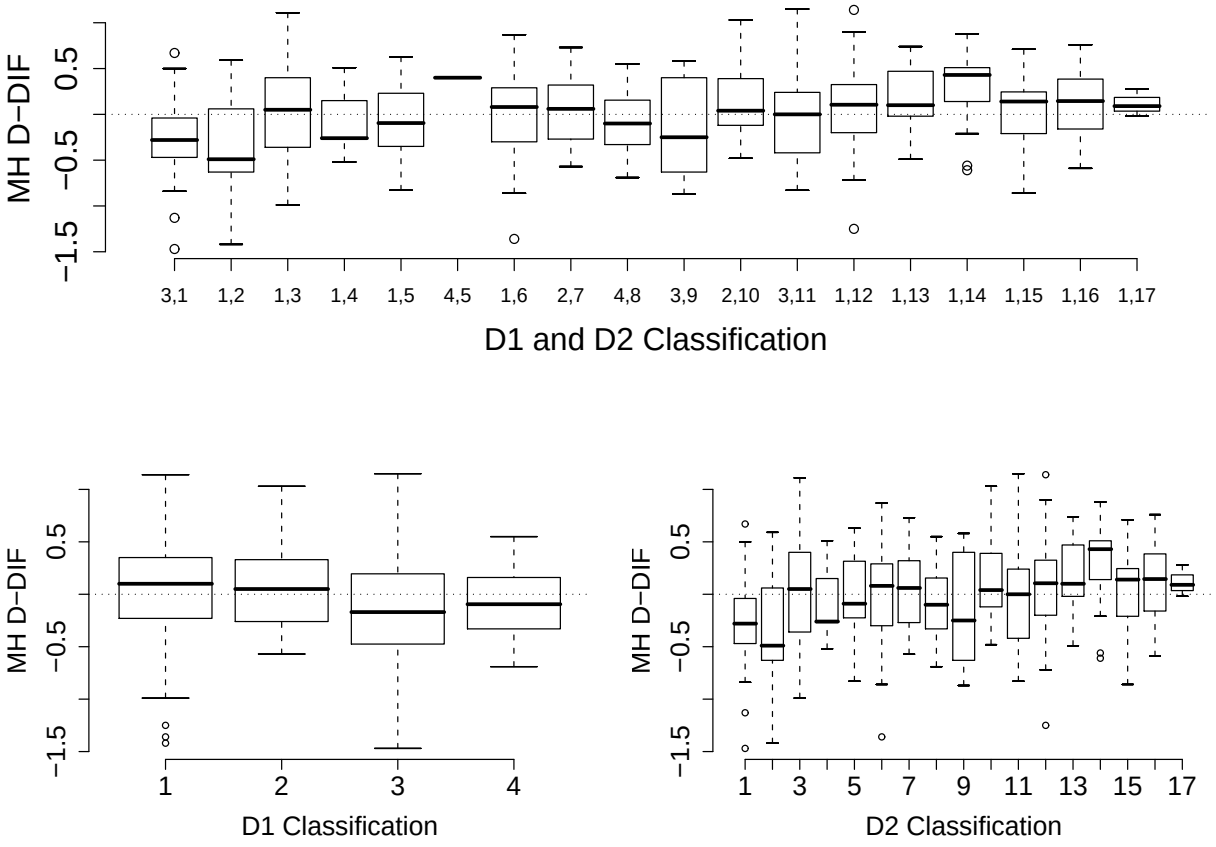

Figure A3. ***Box plots for the MH D-DIF statistics for PPSTW for male-female DIF.***

Note. The dotted line corresponds to an MH D-DIF value of 0.

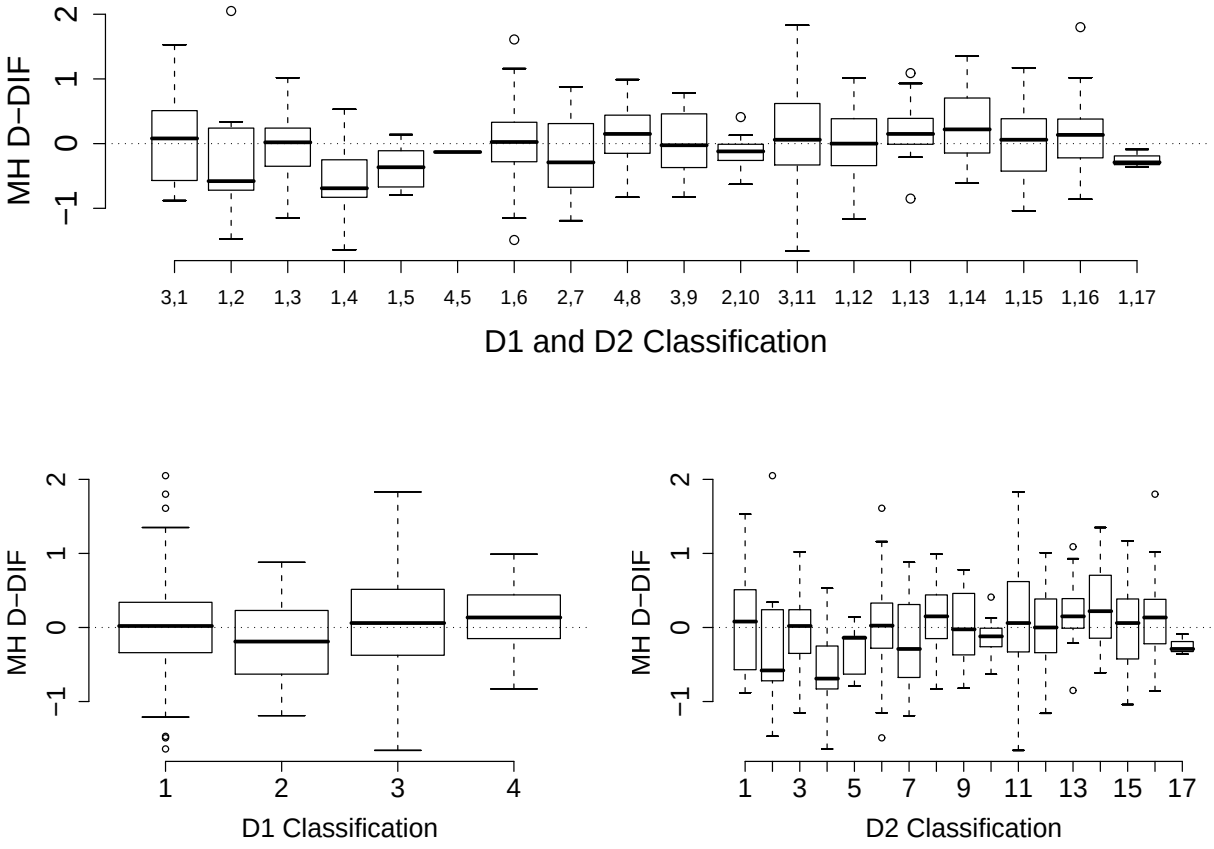

Figure A4. ***Box plots for the MH D-DIF statistics for PPSTW for White-Black DIF.***

Note. The dotted line corresponds to an MH D-DIF value of 0.

Table A3

The Outcome for Male-Female and White-Black DIF for PPSTR

Simul. cond.	Method	Root av. sq. bias	Av. RMSE	Av. % corr. keep	Av. % corr. flag	Av.% corr. all	Av. loss keep	Av loss flag	Av. loss all
MF	C rule	0.04	0.97	98.91	-	98.91	0.14	-	0.14
Data 1	B rule	0.04	0.97	87.99	-	87.99	0.23	-	0.23
165	EB	0.31	0.31	99.96	-	99.96	0.14	-	0.14
50	FB (D1)	0.30	0.30	100.00	-	100.00	0.14	-	0.14
0, 0	FB (1)	0.31	0.31	100.00	-	100.00	0.14	-	0.14
MF	C rule	0.01	0.35	99.84	-	99.84	0.14	-	0.14
Data 1	B rule	0.01	0.35	95.02	-	95.02	0.16	-	0.16
992	EB	0.16	0.25	99.89	-	99.89	0.14	-	0.14
300	FB (D1)	0.16	0.24	99.89	-	99.89	0.14	-	0.14
0, 0	FB (1)	0.16	0.25	99.89	-	99.89	0.14	-	0.14
MF	C rule	0.06	1.03	98.91	5.65	94.00	0.15	1.35	0.21
Data 2	B rule	0.06	1.03	88.65	27.50	85.43	0.23	1.27	0.29
167	EB	0.39	0.35	99.91	0.00	94.65	0.14	1.37	0.21
50	FB (D1)	0.39	0.35	100.00	0.00	94.74	0.14	1.37	0.21
0, 2	FB (1)	0.39	0.35	100.00	0.00	94.74	0.14	1.37	0.21
WB	C rule	0.06	0.88	98.66	5.80	96.22	0.26	1.09	0.28
Data 1	B rule	0.06	0.88	85.05	34.20	83.71	0.35	1.06	0.37
50	EB	0.37	0.42	99.74	0.30	97.12	0.25	1.09	0.27
298	FB (D2)	0.37	0.41	98.91	21.20	96.86	0.26	1.07	0.28
0, 1	FB (1)	0.37	0.42	99.84	0.30	97.22	0.25	1.09	0.27
WB	C rule	0.03	0.31	99.83	4.80	97.33	0.25	1.09	0.27
Data 1	B rule	0.03	0.31	92.23	52.20	91.18	0.29	1.05	0.31
300	EB	0.15	0.27	98.62	19.00	96.53	0.26	1.08	0.28
1792	FB (D2)	0.15	0.26	98.70	38.20	97.11	0.26	1.06	0.28
0, 1	FB (1)	0.15	0.27	98.60	19.00	96.50	0.26	1.08	0.28
WB	C rule	0.08	0.96	98.70	5.87	84.04	0.26	1.40	0.44
Data 2	B rule	0.08	0.96	85.70	28.10	76.60	0.35	1.29	0.49
50	EB	0.51	0.50	99.66	0.68	84.03	0.25	1.44	0.44
277	FB (D2)	0.49	0.48	98.80	9.23	84.66	0.26	1.34	0.43
0, 6	FB (1)	0.51	0.50	99.90	0.68	84.24	0.25	1.44	0.44

Note. The Simul. cond. column gives the type of DIF analysis done (i.e., MF for male-female or WB for White-Black), data set (1 or 2), number of examinees in the focal and reference groups, number of true C DIF items, and number of true B DIF items according to the ETS DIF rule.