Research Report

A Comparison of Methods for Estimating Conditional Item Score Differences in Differential Item Functioning (DIF) Assessments

Tim Moses, Jing Miao, and Neil Dorans

June 2010

ETS RR-10-15

A Comparison of Methods for Estimating Conditional Item Score Differences in Differential Item Functioning (DIF) Assessments

Tim Moses, Jing Miao, and Neil Dorans
ETS, Princeton, New Jersey

June 2010

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: Daniel Eignor

Technical Reviewers: Longjuan Liang and Raymond Mapuranga

Abstract

This study compared the accuracies of four differential item functioning (DIF) estimation methods, where each method makes use of only one of the following: raw data, logistic regression, loglinear models, or kernel smoothing. The major focus was on the estimation strategies' potential for estimating score-level, conditional DIF. A secondary focus was on assessing the accuracy of strategies' overall DIF effect sizes and statistical significance tests. A real data simulation was used to evaluate the estimation strategies with 6 items representing DIF and No DIF situations, and with 4 sample size combinations for the reference and focal group data. Results showed that the logistic regression estimation strategy was the most highly recommended strategy in terms of the bias and variability of its estimates and the power of its statistical significance test. The loglinear models strategy had flexibility advantages, but these advantages only offset the greater variability of its estimates and its reduced statistical power when sample sizes were large. The kernel smoothing estimation strategy was the least accurate of the considered strategies due to estimation problems when the reference and focal groups differed in overall ability.

Key words: DIF, kernel smoothing, loglinear models, logistic regression

Table of Contents

List of Tables

Page

List of Figures

Page

While the psychometric literature has defined differential item functioning (DIF) as a performance difference between examinee groups at one level of ability (Dorans & Holland, 1993; Lord, 1980; Shepard, 1982), considerable research has focused on developing and comparing DIF detection methods that summarize DIF across a total range of ability (Dorans & Kulick, 1986; Holland & Thayer, 1988; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Roussos & Stout, 1996; Shealy & Stout, 1993; Swaminathan & Rogers, 1990; Zumbo, 1999; Zwick, Thayer, & Lewis, 2000). This work usually focuses on overall statistical significance tests of summary DIF indexes and, to a lesser extent, on the use of summary DIF indexes as overall effect sizes. Due to the potential of all summary measures to oversummarize in special circumstances (to be described below), effect sizes and significance tests of overall DIF may benefit by being supplemented with assessments of conditional, ability-level DIF. The purpose of this study was to compare the accuracies of four DIF estimation strategies for estimating conditional DIF (raw data, logistic regression, loglinear models, and kernel smoothing).

Assessing Differential Item Functioning (DIF)

The assessment of DIF is a determination of whether a studied item, *Y*, performs differently for reference examinees, *R*, and focal examinees, *F*, conditioned on the *M* levels of a variable that measures reference and focal examinees' overall ability, X_m. In this study, *Y* is dichotomously scored. X_m denotes an observed test score that excludes *Y* and all items containing extensive DIF, or *C-DIF* (Dorans & Holland, 1993).

The extent of item *Y*'s DIF can be assessed by determining if the reference and focal conditional expected scores differ for any of the *M* levels of X_m,

$$\text{Conditional DIF}_{\text{m}} = E(Y_{Fm}) - E(Y_{Rm}) \neq 0, \quad m = 1, ..., M \quad . \tag{1}$$

In typical DIF assessments, the *M* differences in (1) are summarized rather than individually evaluated. One common DIF summary measure is a focal-weighted average of (1)'s C-DIF estimates,

$$\sum_m \frac{n_{Fm}}{\sum_m n_{Fm}} \left(E(Y_{Fm}) - E(Y_{Rm}) \right) = \sum_m \frac{n_{Fm}}{\sum_m n_{Fm}} \left(\text{Conditional DIF}_{\text{m}} \right) \quad , \tag{2}$$

where n_{Fm} denotes the number of focal examinees at X_m. The DIF summary statistic in (2) is referred to as a standardized expected score difference (i.e., standardized E-Dif; Dorans & Schmitt, 1993). The standardized E-Dif is used mostly as an effect size measure of overall DIF, but since an estimate of its standard error is available (Dorans & Holland, 1993), it can also be used as a statistical significance test.

Potential difficulties with the standardized E-Dif measure are that it can downplay DIF in easy and hard items (Dorans & Holland, 1993, p. 59) or in items exhibiting large degrees of nonuniform DIF. In addition, the standardized E-Difs most frequently described weighting strategy, $\frac{n_{Fm}}{\sum_m n_{Fm}}$, may not be the most appropriate for particular purposes, such as evaluating DIF in the proximity of potential cut scores. To address these issues, it can be useful to supplement overall effect size and significance test DIF assessment by also assessing the *M* differences in (1) with respect to magnitude and with respect to the *M* conditional standard errors,

$$SE(E(Y_{Fm}) - E(Y_{Rm})) = \sqrt{Var(E(Y_{Fm})) + Var(E(Y_{Rm}))}, \quad (3)$$

where the $Var(E(Y))$ terms are the estimated variances of the expected item scores, $E(Y)$. The assessment of (1) and (3) using different DIF estimation strategies is the major focus of this study.

Differential Item Functioning (DIF) Estimation Strategies

This section summarizes the raw, logistic regression, loglinear models, and kernel smoothing DIF estimation strategies of interest in a general overview and as applied to a specific DIF example. Specific details are given in Appendixes A, B, C, and D for how each estimation strategy can be used in (1), (2), and (3) for estimating conditional DIF, conditional standard errors, and the standardized E-Dif measure, and for overall statistical significance tests.

The use of raw data for estimating conditional means and variances in DIF (Appendix A) has been described for estimating the standardized E-Dif measure, for plots of conditional differences (Dorans & Holland, 1993; Dorans & Kulick, 1986), and also for overall statistical significance tests in the related simultaneous item bias test (SIBTEST) framework (Shealy &

Stout, 1993). Raw data offers the most direct approach to DIF estimation and has the least potential for model misspecification error of the strategies considered in this study. The use of raw data produces conditional DIF estimates that are relatively unstable in terms of sampling variability, a feature that could make the estimates less useful than those based on other strategies.

The application of logistic regression procedures to DIF assessment (French & Miller, 1996; Jodoin & Gierl, 2001; Kristjansson et al., 2005; Swaminathan & Rogers, 1990) involves predicting the probability of a correct response on *Y* using logistic curves based on X_m, membership in the reference or focal group, and the interaction of group membership and X_m (Appendix B). Logistic regression has been studied as an overall significance test and has received attention for its estimates of conditional DIF (French & Miller, 1996) and effect sizes (Jodoin & Gierl, 2001; Zumbo, 1999). As a significance test, logistic regression has been shown to be a powerful test relative to other strategies (Swaminathan & Rogers, 1990), especially for detecting levels of DIF that are not the same at each level of X_m (i.e., nonuniform DIF). The accuracy of logistic regression's conditional DIF estimates is less clear, as its imposition of logistic curves is the strongest of assumptions made on the data of all the DIF strategies considered in this study, perhaps increasing its potential for biased estimation (Hanson & Feinstein, 1995; Ramsay, 1991).

The polynomial loglinear models assessed in this study were proposed by Hanson and Feinstein (1995). This estimation strategy is based on identifying DIF in terms of differences in four discrete frequency distributions of X_m: the two frequency distributions of the reference group that gets *Y* correct and incorrect, and the two frequency distributions of the focal group that gets *Y* correct and incorrect (Appendix C). Polynomial loglinear models, one of many loglinear modeling proposals for assessing DIF, are iterative and more flexible versions of Mantel-Haenszel (Holland & Thayer, 1988), are more parsimonious than the "saturated" loglinear models described in Mellenbergh (1982), and have an observed score focus rather than other Rasch-focused loglinear models (Kelderman, 1984). Hanson and Feinstein provided demonstrations of the use of polynomial loglinear models for overall significance tests and for conditional DIF estimates. Conditional standard errors were not described. The Hanson and

Feinstein study demonstrated that loglinear models are more flexible and make fewer impositions on the data than logit models (such as logistic regression models).

A final approach that is considered for assessing overall and conditional DIF is based on kernel smoothing (Ramsay, 1991). Kernel smoothing employs weighted averaging to reduce fluctuations in raw data estimates (Appendix D). This study employs kernel smoothing to separately smooth the raw focal and reference $E(Y_m)$'s, an approach that is routinely used at ETS to assess conditional DIF and also to assess items' nonparametric response curves. This version of kernel smoothing differs from prior versions employed in studies of kernel smoothing applications to DIF that are computationally intensive, nonparametric-IRT-based procedures (Douglas, Stout, & DiBello, 1996; Gierl & Bolt, 2001; Lyu, Dorans, & Ramsay, 1995; Ramsay, 2000).

Example. An example is presented to illustrate the distinguishing features of the four DIF estimation strategies. This example is based on the population data of one of the DIF items featured in this simulation study: the Science1 item. This item was flagged as a conditional DIF item favoring the male reference group (N = 34,336) as compared to the female focal group (N = 18,560). More specific information about the DIF context of this item is described in this study's section, Raw Population Data and Their Population Differential Item Functioning (DIF) Statistics.

The standardized E-Dif values and overall significance tests based on the four DIF estimation strategies of interest are presented in Table 1. The standardized E-Dif values based on raw data, logistic regression, and loglinear models are identical when rounded to their first three decimal places (-0.140). The standardized E-Dif value based on kernel smoothing is somewhat different from those of the other three estimation strategies (-0.148). All four estimation strategies indicate statistically significant overall DIF.

The conditional DIF and +/- 2 estimated standard error bands for the four estimation strategies are presented in Figures 1 to 4. The figures suggest that the Science1 item's DIF is nonuniform (i.e., the level of DIF is not the same across the score levels of X_m). Specifically, DIF is shown to be large and statistically significant for the low-to-middle scores of X_m but close to zero (i.e., no DIF) and possibly statistically insignificant for the higher scores of X_m.

These nonuniform, X_m-varying conditional DIF estimates are missed when the focus is only on the overall standardized E-Dif values and significance tests (Table 1).

Table 1

Comparing the Four Differential Item Functioning (DIF) Estimation Strategies' Overall DIF Assessments in the Study's Population Data (Science1 Item)

Method	Standardized E-Dif	Significance test statistic
Raw data	-0.140	z = -31.03*
Logistic regression	-0.140	χ^2 = 1,146.62* (df = 2)
Loglinear models	-0.140	χ^2 = 1,167.89* (df = 5)
Kernel smoothing	-0.148	z = -33.78*

* p < .05.

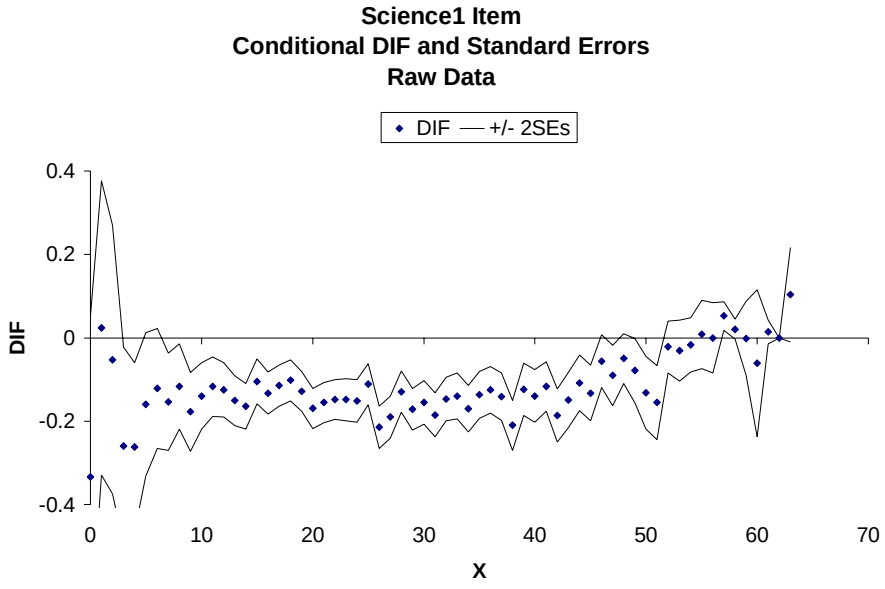


Figure 1. **Science1 item: Raw data for conditional differential item functioning (DIF) and standard errors (SE).**

Figures 1 to 4 illustrate how the four DIF estimation strategies differ: The conditional DIF estimates based on the raw data exhibit large fluctuations and relatively wide standard error bands (Figure 1), while the logistic regression method has narrow standard error bands and conditional DIF estimates that disagree with the raw data's no DIF suggestion at the highest X_m scores (Figure 2 vs. Figure 1). The loglinear model (Figure 3) and kernel smoothing (Figure 4) estimation strategies appear to reflect the trends in Figure 1's raw data conditional DIF estimates more closely than the logistic regression method, with the standard error bands based on the loglinear model being wider than those of the kernel smoothing method at the lowest X_m scores.

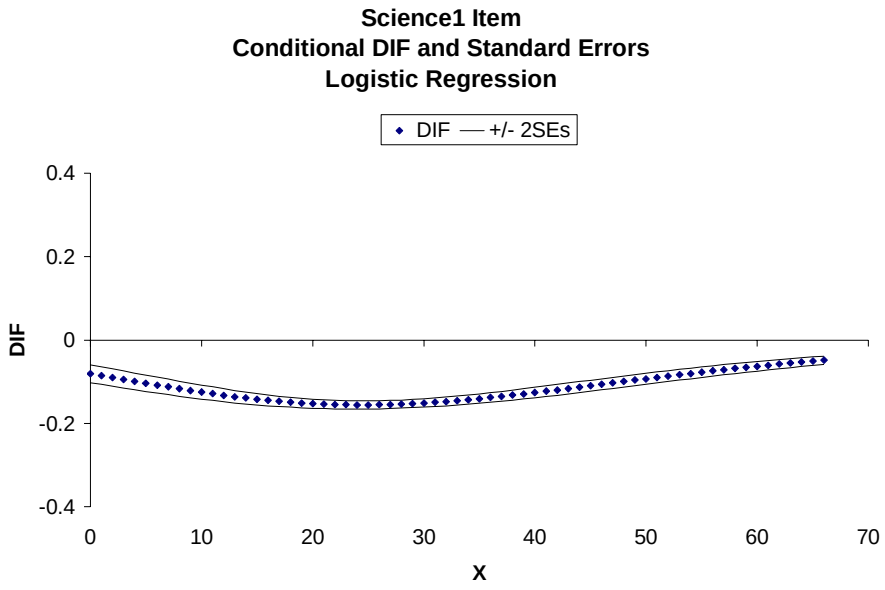


Figure 2. **Science1 item: Logistic regression for conditional differential item functioning (DIF) and standard errors (SE).**

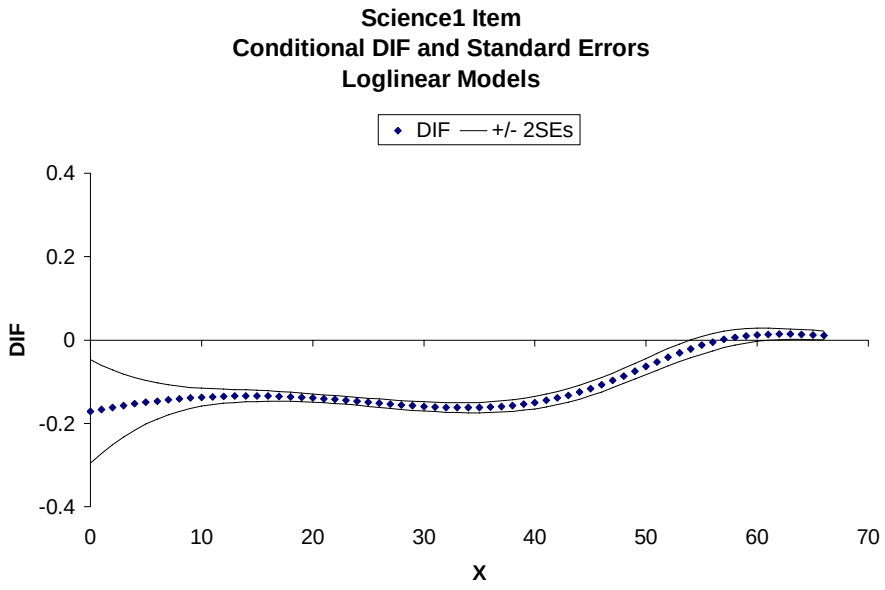


Figure 3. **Science1 item: Loglinear models for conditional differential item functioning (DIF) and standard errors (SE).**

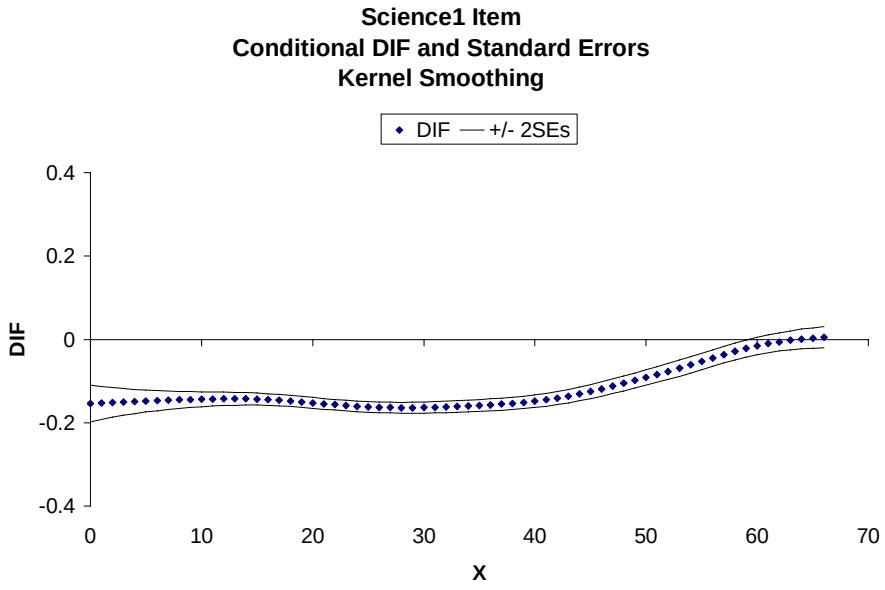


Figure 4. **Science1 item: Kernel smoothing for conditional differential item functioning (DIF) and standard errors (SE).**

This Differential Item Functioning (DIF) Study

This DIF study is different from prior DIF studies in that the major focus is on the accuracy of estimation strategies' conditional DIF and conditional standard error estimates, with somewhat less emphasis on the accuracy of their overall statistical significance tests and overall effect sizes (i.e., standardized E-Dif values; Dorans & Kulick, 1986). As implied in the reviews of the DIF estimation strategies of interest (raw data, logistic regression, loglinear models, and kernel smoothing), much of the prior research has not compared many of these estimation strategies directly to each other and with respect to this study's conditional DIF focus. What studies have been done suggest the following findings from comparisons of the four DIF estimation strategies:

- The estimation strategies may differ more with respect to their conditional DIF estimates than with respect to their ability to estimate the same summary DIF measure, the standardized E-Dif. This suggestion is based on prior studies that assessed the use of various modeling strategies for smoothing raw DIF estimates, which have shown that smoothing conditional DIF estimates and then aggregating these estimates into overall DIF measures does not improve overall DIF measures relative to simply using the raw data (Douglas et al., 1996; Puhan, Moses, Yu, & Dorans, 2007).
- An important issue in comparing the DIF estimation strategies is assessing them in terms of their tradeoff of flexibility to fit a range of conditional DIF curves versus statistical power to detect DIF. Specifically, the logistic regression strategy's imposition of logistic functions onto the sample data is a less flexible and less data-adaptive estimation approach than loglinear models (Hanson & Feinstein, 1995), kernel smoothing (Ramsay, 1991), and raw data. Logistic regression's reduced flexibility could result not only in reduced estimation accuracy for certain DIF situations, but also in increased statistical power because its overall chi-square tests are based on fewer degrees of freedom than that of the loglinear models estimation strategy (Appendixes B and C) and perhaps because its use of simpler modeling parameterizations produce smaller standard errors for conditional DIF estimates.
- The kernel smoothing estimation strategy has its own distinguishing features that need to be compared with those of the other strategies. The described example showed that the conditional DIF and standardized E-Dif based on kernel smoothing differed from those of the other estimation strategies. The use of kernel smoothing as an overall statistical significance

test is an additional interest, as this issue has received little attention in prior studies and has not resulted in an extremely accurate significance test (Douglas et al., 1996).

Method

The raw data, logistic regression, loglinear modeling, and kernel smoothing DIF estimation strategies were compared in several simulations. Populations for DIF items were obtained from large-volume test data, and the DIF statistics computed from the raw population data were used as population DIF statistics. From these populations, sample datasets were randomly drawn at specific reference and focal group sample sizes. Conditional and overall DIF were assessed using each of the four strategies in each of the sample datasets. The accuracies of the estimation strategies were studied by averaging their results over 400 replications of sample datasets and comparing the averages to the population DIF statistics computed in the raw population data.

Raw Population Data and Their Population Differential Item Functioning (DIF) Statistics

The study used test data from two large-scale achievement tests as the populations. These populations are comprised of test data used to conduct actual DIF analyses, making these populations especially useful for realistic evaluations that are relevant for practice. Six conditional DIF items were found, three from a 69-item science test and three from an 80-item history test. The DIF was based on males and females, a comparison that resulted in large reference and focal populations. The science test data consisted of 52,896 examinees, with 34,336 examinees in the reference group (i.e., male) and 18,560 examinees in the focal group (i.e., female). The history test data consisted of 325,250 examinees, with 147,737 examinees in the reference group (i.e., male) and 177,513 examinees in the focal group (i.e., female).

Table 2 presents the population statistics of the six items, including their average item scores, point-biserial correlations with the matching variable, and the standardized average reference versus focal difference on the matching variable. Table 2 also shows the items' standardized E-Dif values calculated from the raw population data (used as population DIF statistics in this study). Table 2's summary of the six items shows that these items vary in their DIF-relevant characteristics, including different levels of reference versus focal abilities on the matching variable (the science vs. history items), easier and more difficult studied items (Science3 vs. the other five items), varied correlations with the matching variable, varied

magnitudes of DIF (Science1 and Science2 vs. Science3 and History2), and DIF situations where the reference group is favored (Science1 and Science2) and other DIF situations where the focal group is favored (Science3, History1, History2, and History3).

Sample Sizes

Random samples of the reference and focal data were drawn from the population data in reference/focal sizes of 2,000/2,000, 2,000/700, 700/700, and 700/200.

Simulations

The simulations were conducted to assess the raw, logistic regression, loglinear models, and kernel smoothing DIF estimation strategies with respect to their estimation of six different items and four reference and focal sample size combinations. For each of the 6 x 4 = 24 combinations of DIF item and sample size, 400 datasets were randomly drawn from the population data. In each of these sample datasets, the four DIF estimation strategies were used to estimate the conditional DIF in the raw population data across all *M* levels of matching variable X_m (1), to estimate the *M* conditional standard errors of the conditional DIF estimates (3), to conduct significance tests of overall DIF (Appendixes A, B, C, and D), and to estimate the overall standardized E-Dif measure (2) in the raw population data.

Table 2

Summary of the Raw Population Data for the Six Studied Items (Y)

Subject & item (*Y*)	Average item score on *Y* in the combined focal and reference data	Point-biserial correlation between *X* and *Y* in the combined focal and reference data	Average standardized difference on X, (focal-reference)	Standardized E-Dif of *Y* based on raw data
Science1	0.69	0.32	-0.41	-0.14
Science2	0.75	0.42	-0.41	-0.12
Science3	0.23	0.44	-0.41	0.07
History1	0.77	0.38	-0.26	0.10
History2	0.92	0.27	-0.26	0.08
History3	0.76	0.40	-0.26	0.10

The study also considered 24 additional no DIF conditions for the six items and four sample sizes. For the no DIF conditions, the studied item's conditional expected scores computed in the combined population reference and focal data were used as population parameters for randomly generating the reference and focal studied item responses in each of the sample datasets. The data generation for the no DIF conditions is illustrated in following three bullets:

- For the Science1 item, the expected score of the combined population reference and population focal data at X_m = 5 was 0.358. For the simulation of no DIF in the Science1 item, the Science1 item scores for the reference and focal data at X_m = 5 were created by randomly drawing values of either 0 or 1, where the probability of drawing a score of 1 at X_m = 5 was 0.358. The result of this generation of Science1 item scores was that the expected (population) Science1 item score at X_m = 5 was the same (no DIF) in the reference and focal sample data, 0.358.
- For the Science3 item, the expected score of the combined population reference and population focal data at X_m = 11 was 0.054. For the simulation of no DIF in the Science3 item, the Science3 item scores for the reference and focal data at X_m = 11 were created by randomly drawing values of either 0 or 1, where the probability of drawing a score of 1 at X_m = 11 was 0.054. The result of this generation of Science3 item scores was that the expected (population) Science3 item score at X_m = 11 was the same (no DIF) in the reference and focal sample data, 0.054.
- For the History2 item, the expected score of the combined population reference and population focal data at X_m = 27 was 0.849. For the simulation of no DIF in the History2 item, the History2 item scores for the reference and focal data at X_m = 27 were created by randomly drawing values of either 0 or 1, where the probability of drawing a score of 1 at X_m = 27 was 0.849. The result of this generation of History2 item scores was that the expected (population) History2 item score at X_m = 27 was the same (no DIF) in the reference and focal sample data, 0.849.

For all of the no DIF conditions, the scores for all X_m values of all six items were generated in the same manner as what was described in the previous three bullets. These scores resulted in reference and focal data where the expected (population) DIF was zero for all X_m values (1) and also zero when aggregated across all X_m values (2). This generation of no DIF made it possible for the DIF strategies to be assessed in no DIF conditions that preserved the overall characteristics of the studied items (i.e., difficulty, point-biserial correlations) and matching variables (score ranges, overall reference, and focal ability differences).

For each of the 48 total conditions (4 sample sizes X 6 items X DIF vs. no DIF = 48), the accuracies of the four DIF estimation strategies' conditional DIF and conditional standard error estimates, overall significance tests, and standardized E-Dif measures were assessed as averaged across the 400 replicated datasets and compared with the values computed in the raw population data.

Accuracy measures. To evaluate the accuracy of the conditional DIF estimates for each of the study's 48 conditions (six studied items, four sample size combinations and DIF vs. no DIF conditions), measures were computed from the mean squared error (*MSE*) calculated at each of the *M* levels of the matching variable, X_m,

$$\begin{aligned} MSE_m &= \frac{1}{400}\sum_{replication}\left(\hat{\theta}_{m,replication}-\theta_m\right)^2 \\ &= \frac{1}{400}\sum_{replication}\left[\left(\bar{\theta}_m-\theta_m\right)^2+\left(\hat{\theta}_{m,replication}-\bar{\theta}_m\right)^2\right] \\ &= Bias_m^2+Variance_m \end{aligned} \quad , \qquad (4)$$

where *replication* indicates one of the 400 random datasets drawn from one of the population distributions at one of the four sample size combinations, $\hat{\theta}_{m,replication}$ is the estimated conditional DIF estimate in one of the 400 datasets at X_m, $\bar{\theta}_m$ is the average of the 400 sample datasets' conditional DIF estimates at X_m, and θ_m is the conditional DIF estimate computed in the raw population data at X_m.

The square roots of the squared conditional squared bias and variance in (4) were taken and averaged with respect to the *M* score levels of X_m to form average absolute conditional bias and average conditional standard deviations,

$$\text{Avg. Abs. Conditional Bias} = \frac{1}{M}\sum_{m}\sqrt{Bias_m^2}\,, \tag{5}$$

$$\text{Avg. Conditional SD} = \frac{1}{M}\sum_{\text{m}}\sqrt{Variance_m} = \frac{1}{M}\sum_{\text{m}} SD_m\,. \tag{6}$$

To assess the extent to which strategies' estimated conditional standard errors approximated their estimates' actual variability (i.e., the SD_m's in (6)), a measure similar to (5) was used,

$$\text{Avg. Abs. Conditional SE Inaccuracy} = \left(\frac{1}{M}\right)\frac{\sum_{m}\sqrt{\left(\widehat{SE}_m - SD_m\right)^2}}{\text{Avg. Conditional SD}}\,, \tag{7}$$

where $\widehat{SE}_m$ is the average of a DIF strategy's estimated conditional standard errors across the 400 replications of an item and sample size combination.

Alternative summary measures to (5), (6), and (7) would be to average the X_m-level squared differences or the signed differences rather than the absolute differences, and/or weight the X_m-level results by a population distribution. The X_m-level averaging was done on the absolute differences because it oriented the averaging directly on the conditional DIF and standard error quantities of interest rather than on the squared values. The averaging of absolute differences was desirable also because it produced summaries that were not influenced by the cancellation of positive and negative differences. The nonweighting in (5), (6), and (7) was used because, in practice, conditional DIF results would potentially be evaluated at score levels not necessarily based on where the most data are found. Preliminary evaluations of the results showed that the conclusions would not be dramatically altered by using alternative versions of (5), (6), and (7), but they would also not be identical to the reported results. Plots of the strategies' estimation results were also created to supplement the summary measures. These plots

depicted the biases of the conditional DIF ($\bar{\theta}_m - \theta_m$), and the size and accuracies of the standard error estimates ($\widehat{SE}_m$ vs. SD_m) for specific item and sample size combinations of interest.

To evaluate the DIF strategies' accuracy in terms of the standardized E-Dif measure, accuracy measures were created as the square roots of the squared bias (standardized E-Dif absolute bias) and variance (standardized E-Dif SD) parts of its own MSE,

$$\text{Standardized E-Dif Absolute Bias} = \sqrt{\left(\bar{\theta} - \theta\right)^2} = \sqrt{Bias^2}, \tag{8}$$

$$\text{Standardized E-Dif SD} = \sqrt{\frac{1}{400}\sum_{replication}\left(\hat{\theta}_{replication} - \bar{\theta}\right)^2} = \sqrt{Variance}. \tag{9}$$

where $\hat{\theta}_{replication}$ is the estimated standardized E-Dif value in one of the 400 datasets, $\bar{\theta}$ is the average of the 400 sample datasets' standardized E-Dif values, and θ is the raw data standardized E-Dif value computed in the population data.

The accuracy of the DIF estimation strategies' overall statistical significance tests was also assessed. For this evaluation, a rate was calculated for how often each estimation strategy indicated that DIF was statistically significant across the 400 replications of an item and sample size condition. When the studied item responses were drawn from the actual male and female population data, these rates were power rates (i.e., the rate at which the DIF estimation strategies correctly indicated DIF when DIF was in the population). When studied item responses for the male and female samples were randomly generated from a common set of conditional expected scores, these rates were Type I error rates (i.e., the rate at which the DIF estimation strategies incorrectly indicated DIF when DIF was not in the population). The superior strategy in terms of statistical significance tests was the one that had the largest power rate while staying sufficiently close to the desired 0.05 Type I error rate, where sufficient was defined as within a range of 0.025 to 0.075. This range is known as Bradley's (1978) liberal criterion of robustness and is commonly used to evaluate statistical strategies' Type I error rates (e.g., Keselman, Wilcox, Othman, & Fradette, 2002).

Results

Differential Item Functioning (DIF) Estimation Strategies' Conditional DIF and Standard Error Results

The results of DIF estimation strategies' conditional DIF and standard error estimates are summarized by studied item (measures are averaged across the 4X2=8 combinations of sample size and DIF vs. no DIF; Table 3), by sample size (measures are averaged across the 6X2=12 combinations of studied item and DIF vs. no DIF; Table 4) and by DIF versus no DIF (measures are averaged across the 6X4=24 combinations of studied item and sample size; Table 5). Each of these tables compares the four estimation strategies in terms of the extent to which their conditional DIF estimates systematically deviated from the population conditional DIF values (average. absolute conditional bias, or avg. abs. conditional bias), the variability of their conditional DIF estimates (average conditional standard deviation, or avg. conditional SD), and the accuracy of their conditional standard errors (average absolute conditional standard error inaccuracy, or avg. abs. conditional SE inaccuracy). The values of absolute bias, variability, and standard error accuracy for specific items, sample sizes, and DIF condition are bolded to indicate the best DIF estimation strategy and underlined to indicate the worst DIF estimation strategy.

The DIF estimation strategies produced mixed results in terms of their absolute conditional bias for the items, sample sizes, and DIF conditions. The raw data strategy had the smallest absolute conditional biases for the three science items, while the loglinear models strategy had the smallest values for the History1 item and the logistic regression strategy had the smallest values for the History2 and History3 items. The kernel smoothing strategy had the largest absolute conditional biases for four of the six studied items. In terms of sample sizes, the logistic regression strategy had the smallest absolute conditional bias for the smallest sample size condition considered (700/200), while the raw data strategy had the smallest absolute conditional biases and the kernel smoothing strategy had the largest absolute conditional biases for the three larger sample size conditions (700/700, 2,000/700, and 2,000/2,000). For the no DIF conditions, the logistic regression strategy had the smallest absolute conditional bias and the raw data strategy had the largest absolute conditional bias. For the DIF conditions, the raw data strategy had the smallest absolute conditional bias while the kernel smoothing strategy had the largest absolute conditional bias.

Table 3

The Four Differential Item Functioning (DIF) Estimation Strategies' Conditional DIF and Standard Error (SE) Results by Item

	Avg. abs. conditional bias				Avg. conditional SD				Avg. abs. conditional SE inaccuracy			
Items	Raw	Logistic	Loglinear	Kernel	Raw	Logistic	Loglinear	Kernel	Raw	Logistic	Loglinear	Kernel
Science1	**0.013**	0.023	0.021	0.027	0.214	**0.034**	0.074	0.042	0.413	**0.062**	0.093	0.143
Science2	**0.008**	0.015	0.019	0.026	0.160	**0.025**	0.062	0.032	0.405	**0.048**	0.086	0.131
Science3	**0.011**	0.023	0.019	0.024	0.188	**0.033**	0.070	0.043	0.379	**0.062**	0.106	0.174
History1	0.023	0.017	**0.016**	0.018	0.205	**0.032**	0.083	0.041	0.451	**0.041**	0.084	0.166
History2	0.020	**0.010**	0.013	0.013	0.177	0.036	0.079	**0.033**	0.545	**0.088**	0.082	0.201
History3	0.014	**0.011**	0.014	0.017	0.196	**0.031**	0.076	0.041	0.428	**0.065**	0.082	0.161

Note. The best strategy in terms of absolute bias, standard deviation, and standard error inaccuracy for each item is bolded while the worst strategy is underlined. Avg. abs. = average absolute, SD = standard deviation, SE = standard error.

Table 4

The Four Differential Item Functioning (DIF) Estimation Strategies' Conditional DIF and Standard Error (SE) Results by Sample Size

Sample	Avg. abs. conditional bias				Avg. conditional SD				Avg. abs. conditional SE inaccuracy			
sizes (R/F)	Raw	Logistic	Loglinear	Kernel	Raw	Logistic	Loglinear	Kernel	Raw	Logistic	Loglinear	Kernel
700/200	0.023	**0.017**	0.024	0.023	0.258	**0.049**	0.114	0.057	0.533	**0.049**	0.113	0.268
700/700	0.015	0.016	**0.014**	0.021	0.201	**0.032**	0.075	0.039	0.445	0.084	**0.073**	0.145
2,000/700	**0.012**	0.016	0.017	0.021	0.168	**0.027**	0.063	0.034	0.410	**0.049**	0.094	0.145
2,000/2,000	**0.010**	0.016	0.013	0.019	0.133	**0.019**	0.044	0.025	0.360	**0.062**	0.076	0.092

Note. The best strategy in terms of absolute bias, standard deviation, and standard error inaccuracy for each sample size is bolded while the worst strategy is underlined. Avg. abs. = average absolute, R/F = reference/focal, SD = standard deviation, SE = standard error.

Table 5

The Four Differential Item Functioning (DIF) Estimation Strategies' Conditional DIF and Standard Error (SE) Results by DIF/No DIF Conditions

	Avg. abs. conditional bias				Avg. conditional SD				Avg. abs. conditional SE inaccuracy			
DIF	Raw	Logistic	Loglinear	Kernel	Raw	Logistic	Loglinear	Kernel	Raw	Logistic	Loglinear	Kernel
No	0.015	**0.004**	0.009	0.011	0.191	**0.032**	0.074	0.039	0.440	**0.048**	0.091	0.164
Yes	**0.015**	0.029	0.025	0.031	0.189	**0.032**	0.074	0.039	0.434	**0.074**	0.087	0.160

Note. The best strategy in terms of absolute bias, standard deviation, and standard error inaccuracy for the DIF conditions is bolded while the worst strategy is underlined. Avg. abs. = average absolute, SD = standard deviation, SE = standard error.

The four DIF estimation strategies were fairly consistent in terms of the variability of their conditional DIF estimates (average. conditional standard deviation) across the items (Table 3), sample sizes (Table 4), and DIF versus no DIF conditions (Table 5). The general result was that the most-to-least variable conditional DIF estimates were those based on raw data, loglinear models, kernel smoothing, and logistic regression. The raw data estimates were more than twice as variable as those of the second most variable loglinear models' estimates, which in turn were usually more than twice as variable as those of the least variable logistic regression's estimates.

The four DIF estimation strategies were fairly consistent in terms of the accuracy of their conditional standard error estimates (average absolute conditional standard error inaccuracy) across the items (Table 3), sample sizes (Table 4), and DIF versus no DIF conditions (Table 5). Generally, the most-to-least accurate conditional standard error estimates were those based on logistic regression, loglinear models, kernel smoothing, and raw data.

Plots to further assess the conditional DIF and standard error results. Plots were used to examine the estimation strategies' bias and variability results in detail for a limited number of this study's conditions. These plots focused on the results obtained for the Science1 item, the results of which are representative of the plots produced for the other five items. To consider the biases of the DIF strategies' conditional DIF estimates in the no DIF condition, Figures 5 and 6 plot the strategies' conditional biases, where the studied item had no DIF in the population, and where the reference and focal datasets were drawn at sample sizes of 700/200 (Figure 5) and at sample sizes of 2,000/2,000 (Figure 6). For the small sample size condition shown in Figure 5, the raw data and loglinear models estimation strategies exhibit their largest biases at the highest and lowest scores of X_m, while the kernel smoothing estimation strategy exhibits small but consistently negative biases throughout many of the low to middle scores of X_m. The strategies' conditional biases are generally small when based on large sample sizes (Figure 6), though the raw data biases have fluctuations at the high and low scores of X_m, the loglinear models' biases are largest at the lowest scores of X_m, and the kernel smoothing biases are small but consistently negative for many of the low to middle scores of X_m.

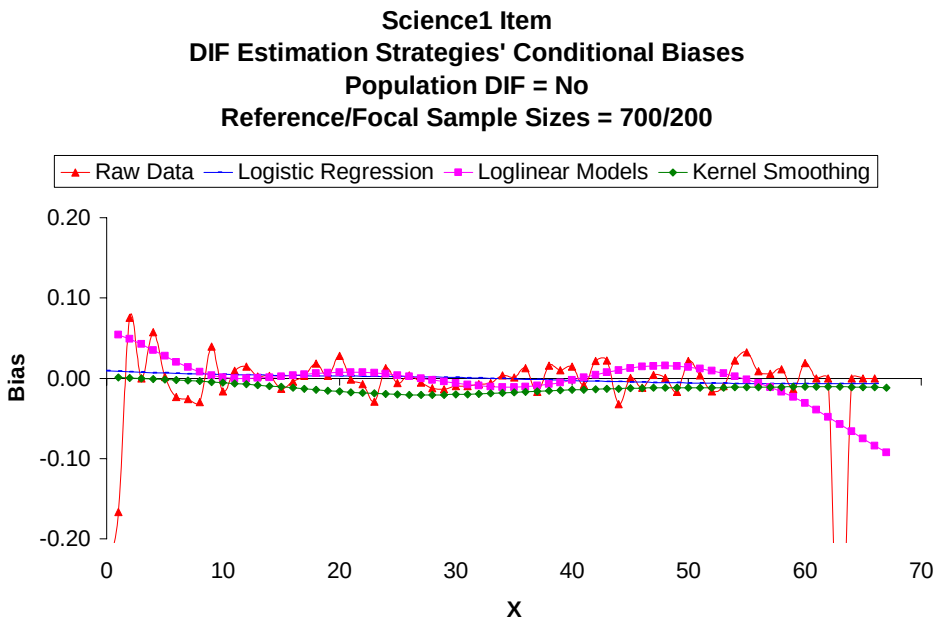


Figure 5. **Science1 item: Differential item functioning (DIF) estimation strategies' conditional biases—population DIF = no, reference/focal sample sizes = 700/200.**

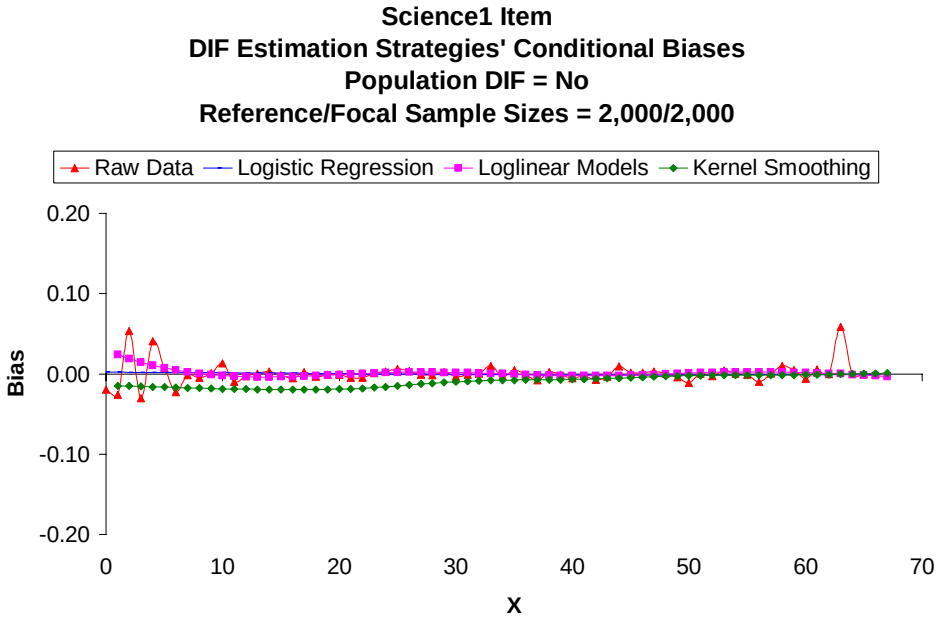
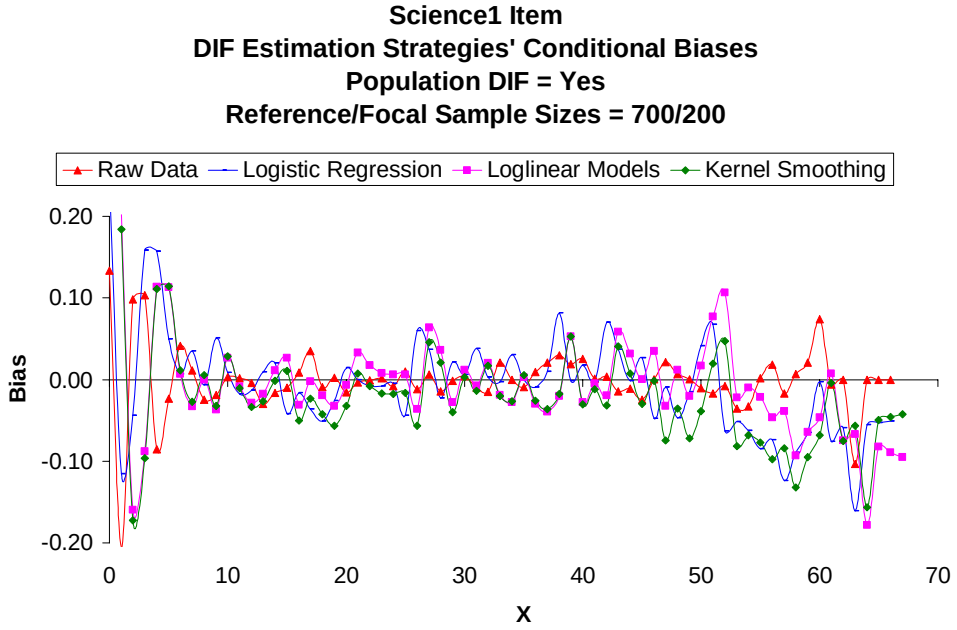


Figure 6. **Science1 item: Differential item functioning (DIF) estimation strategies' conditional biases—population DIF = no, reference/focal sample sizes =2,000/2,000.**

To consider the estimation strategies' biases in conditions where the studied item had DIF in the population, Figures 7 and 8 plot the strategies' conditional biases where the Science1 item had DIF in the population and where the reference and focal datasets were drawn at sample sizes of 700/200 (Figure 7) and at sample sizes of 2,000/2,000 (Figure 8). The bias results in Figures 7 and 8 are very erratic, due in large part to the fluctuations in the population conditional DIF (Figure 1). The major results require close inspection of the figures and show that the raw data estimates are generally less biased than those of the other three DIF estimation strategies, particularly at the highest and lowest scores of X_m. The loglinear models' estimation strategy produced conditional biases that were less accurate than those of the logistic regression estimation strategy for the small sample size condition (Figure 7) and more accurate than those of the logistic regression estimation strategy for the large sample size condition (Figure 8). The kernel smoothing biases deviated from the zero line to a larger extent than the biases based on raw data, loglinear models, and logistic regression estimation strategies.



.

Figure 7. **Science1 item: Differential item functioning (DIF) estimation strategies' conditional biases—population DIF = yes, reference/focal sample sizes =700/200.**

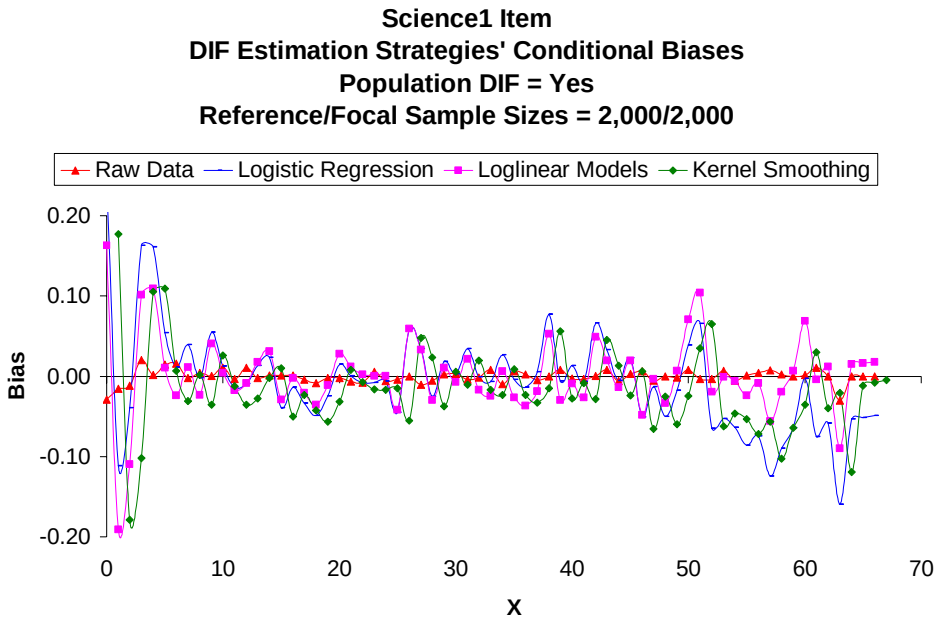


Figure 8. **Science1 item: Differential item functioning (DIF) estimation strategies' conditional biases—population DIF = yes, reference/focal sample sizes = 2,000/2,000.**

To evaluate the DIF strategies' variabilities and the accuracies of their estimated standard errors, Figures 9 to 12 plot the strategies' conditional SD_m and $\widehat{SE}_m$ values obtained from the Science1 item based on reference/focal sample sizes of 700/200 and 2,000/2,000. The major results shown in these plots are that the standard error estimates get smaller and more accurate with larger sample sizes. The standard error estimates based on 700/200 sample sizes using the raw data strategy (Figure 9) are particularly inaccurate in that they underestimate actual variability (i.e., the SD_m's in (6)) for the majority of the X_m scores. The standard error estimates of the logistic regression (Figure 10), loglinear models (Figure 11), and kernel smoothing (Figure 12) estimation strategies are smaller, smoother, and more accurate than those based on raw data.

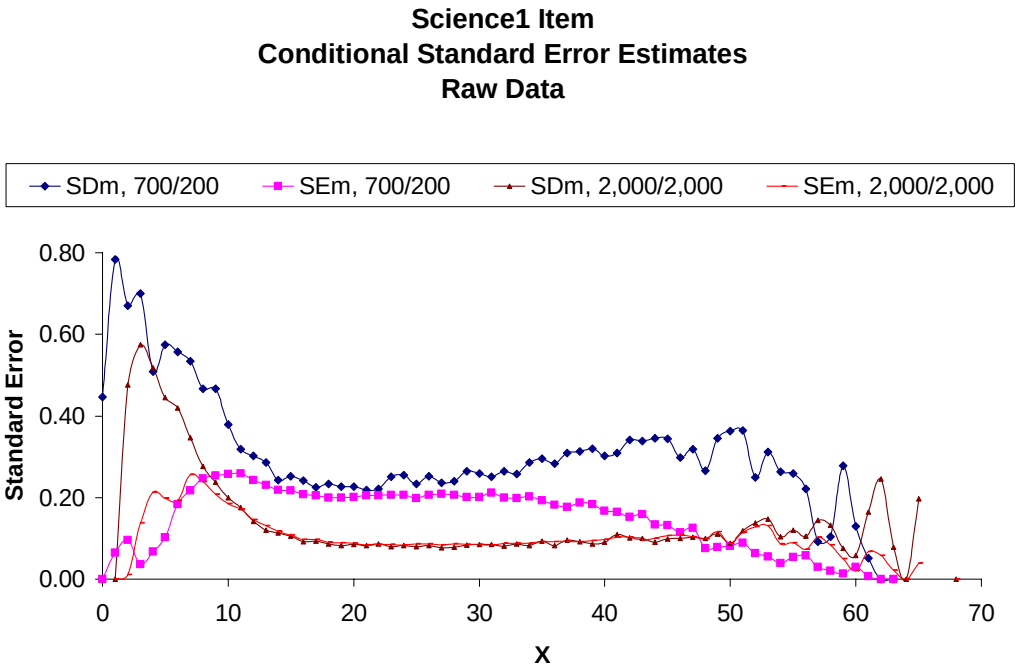


Figure 9. **Science1 item: Raw data for conditional standard error (SE) estimates.**

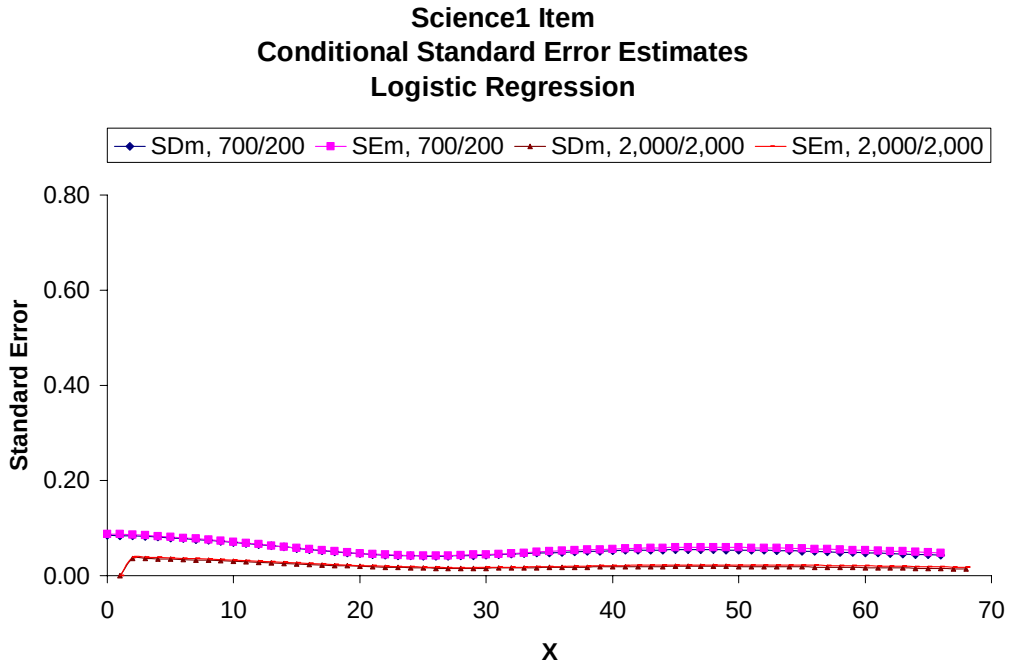


Figure 10. **Science1 item: Logistic regression for conditional standard error (SE) estimates.**

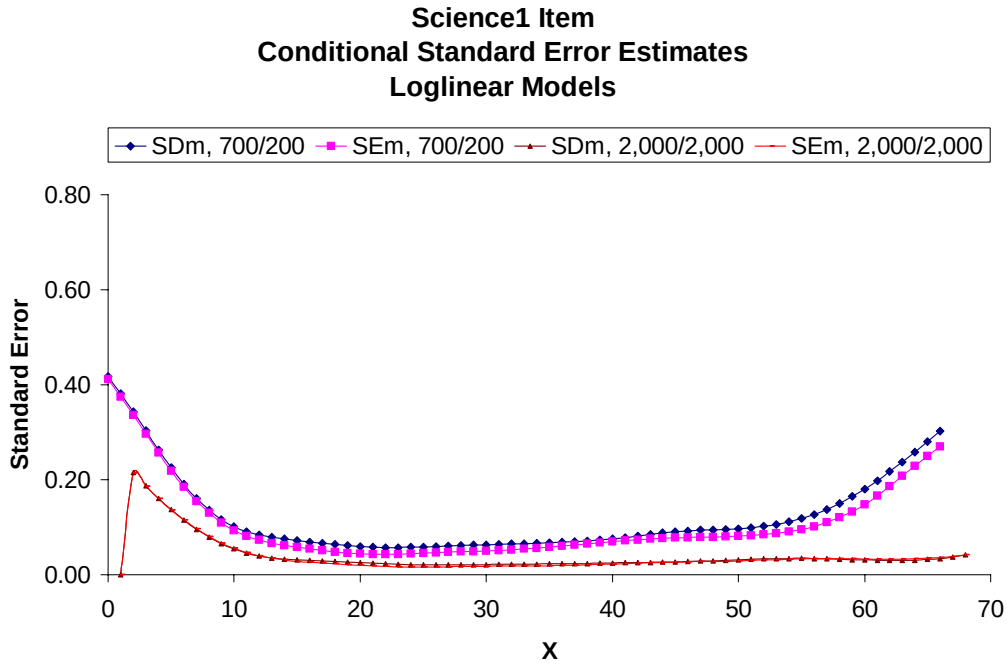


Figure 11. **Science1 item: Loglinear models for conditional standard error (SE) estimates.**

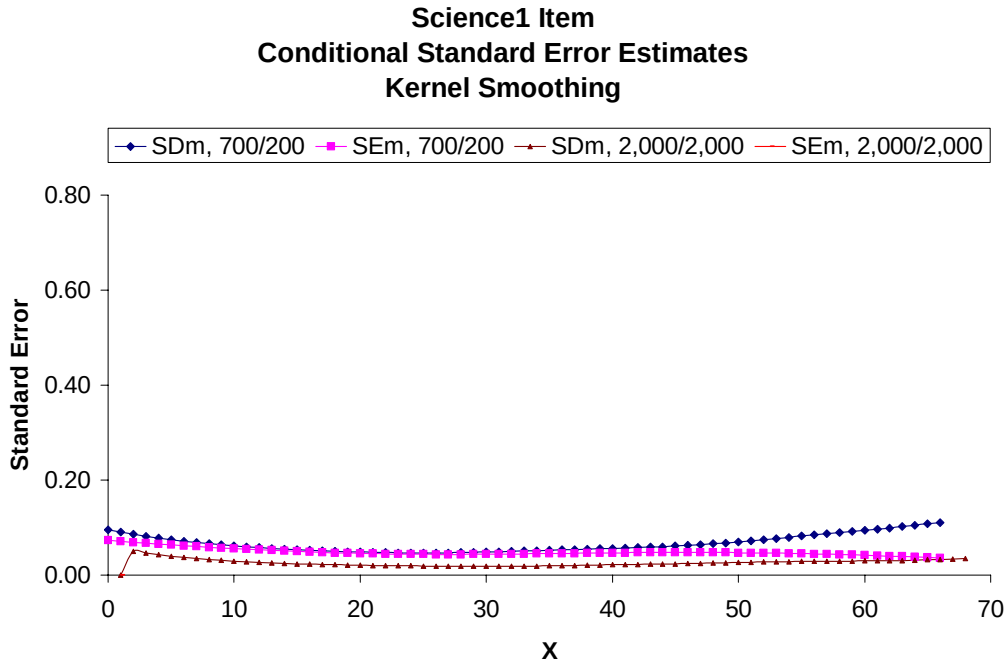


Figure 12. **Science1 item: Kernel smoothing for conditional standard error (SE) estimates.**

Differential Item Functioning (DIF) Estimation Strategies and Standardized E-Dif Estimation

The raw, logistic regression, loglinear models, and kernel smoothing DIF estimation strategies' absolute biases and standard deviations in estimating the standardized E-Dif measure are shown for each item (Table 6), sample size combination (Table 7), and DIF versus no DIF condition (Table 8). The values of absolute bias and variability are bolded to indicate the best DIF estimation strategy and underlined to indicate the worst DIF strategy. In terms of absolute bias, the results show small (0.001) and almost identical absolute biases in the standardized E-Dif values based on raw data, logistic regression, and loglinear models, and larger (greater than 0.010) absolute bias values in the standardized E-Dif values based on kernel smoothing. In terms of standard deviations, the standardized E-Dif values based on raw data exhibited slightly larger (by at most .002) variability than those based on logistic regression, loglinear models, and kernel smoothing.

Table 6

The Four Differential Item Functioning (DIF) Estimation Strategies' Accuracies for the Standardized E-Dif by Item

	Standardized E-Dif absolute bias				Standardized E-Dif SD			
Items	Raw	Logistic	Loglinear	Kernel	Raw	Logistic	Loglinear	Kernel
Science1	**0.001**	**0.001**	**0.001**	0.015	0.026	**0.025**	**0.025**	**0.025**
Science2	**0.001**	**0.001**	**0.001**	0.021	**0.023**	**0.023**	**0.023**	**0.023**
Science3	**0.001**	0.002	**0.001**	0.011	0.019	**0.018**	**0.018**	0.019
History1	**0.001**	**0.001**	**0.001**	0.008	0.022	**0.021**	**0.021**	**0.021**
History2	**0.000**	**0.000**	**0.000**	0.004	0.015	0.015	0.015	**0.014**
History3	**0.001**	**0.001**	**0.001**	0.009	0.022	**0.021**	**0.021**	**0.021**

Note. The best strategy in terms of absolute bias and standard deviation for each item is bolded while the worst strategy is underlined. SD = standard deviation.

Differential Item Functioning (DIF) Strategies' Type I Error and Power Rates

To evaluate the four DIF estimation strategies in terms of the accuracies of their overall statistical significance tests, Table 9 presents their Type I error (no DIF) and power (DIF) rates for the six considered items and Table 10 presents their Type I error rate and power rates for the four reference/focal sample sizes. In terms of Type I error, the raw data, logistic regression, and loglinear models estimation strategies were robust with respect to the 0.025 to 0.075 criterion

range, while the kernel smoothing estimation strategy produced consistently inflated Type I error rates. The estimation strategies could generally be ordered from most to least powerful as kernel smoothing, logistic regression, raw data, and loglinear models. The kernel smoothing estimation strategy's high power rates are not useful due to its inability to sufficiently control Type I error. The loglinear models' estimation strategy had power levels that suffered most in the smallest sample size condition (700/200) and had power levels that were very similar to those of the logistic regression and raw data strategies with the larger sample size conditions.

Table 7

The Four Differential Item Functioning (DIF) Estimation Strategies' Accuracies for the Standardized E-Dif by Sample Size

Sample sizes	Standardized E-Dif absolute bias				Standardized E-Dif SD			
(R/F)	Raw	Logistic	Loglinear	Kernel	Raw	Logistic	Loglinear	Kernel
700/200	**0.001**	**0.001**	**0.001**	0.012	0.033	**0.031**	**0.031**	0.032
700/700	**0.000**	**0.001**	**0.001**	0.012	0.022	**0.021**	**0.021**	**0.021**
2,000/700	**0.001**	**0.001**	**0.001**	0.010	**0.017**	**0.017**	**0.017**	**0.017**
2,000/2,000	**0.001**	**0.001**	**0.001**	0.010	0.013	**0.012**	**0.012**	**0.012**

Note. The best strategy in terms of absolute bias and standard deviation for each sample size is bolded while the worst strategy is underlined. R/F = reference/focal; SD = standard deviation.

Table 8

The Four Differential Item Functioning (DIF) Estimation Strategies' Accuracies for the Standardized E-Dif by DIF/No DIF

	Standardized E-Dif absolute bias				Standardized E-Dif SD			
DIF	Raw	Logistic	Loglinear	Kernel	Raw	Logistic	Loglinear	Kernel
No	**0.001**	**0.001**	**0.001**	0.012	0.021	**0.020**	**0.020**	**0.020**
Yes	**0.001**	**0.001**	**0.001**	0.011	**0.021**	**0.021**	**0.021**	**0.021**

Note. The best strategy in terms absolute bias and standard deviation each DIF condition is bolded while the worst strategy is underlined. SD = standard deviation.

Table 9

The Four Differential Item Functioning (DIF) Estimation Strategies' Type I Error and Power Rates by Item

DIF	Items	Raw	Logistic	Loglinear	Kernel
No	Science1	0.056	0.057	0.069	0.119[a]
(Type I error)	Science2	0.037	0.053	0.045	0.193 [a]
	Science3	0.033	0.054	0.043	0.114 [a]
	History1	0.046	0.054	0.061	0.083 [a]
	History2	0.043	0.056	0.069	0.078 [a]
	History3	0.042	0.055	0.057	0.084 [a]
Yes	Science1	0.985	0.981	<u>0.968</u>	**0.998**
(Power)	Science2	0.959	0.961	<u>0.930</u>	**0.996**
	Science3	0.901	**0.929**	0.898	<u>0.893</u>
	History1	0.959	0.951	<u>0.923</u>	**0.969**
	History2	0.978	0.984	<u>0.961</u>	**0.992**
	History3	0.945	0.949	<u>0.918</u>	**0.960**

Note. The most powerful strategy's power rate is bolded while the least powerful strategy's power rate is underlined.

[a] Nonrobust Type I error rates that are outside the 0.025 to 0.075 range.

Table 10

The Four Differential Item Functioning (DIF) Estimation Strategies' Type I Error and Power Rates by Sample Size

DIF	Sample sizes (R/F)	Raw	Logistic	Loglinear	Kernel
No	700/200	0.052	0.060	0.072	0.103 [a]
(Type I Error)	700/700	0.046	0.053	0.048	0.093 [a]
	2,000/700	0.034	0.047	0.055	0.107 [a]
	2,000/2,000	0.040	0.059	0.054	0.144 [a]
Yes	700/200	0.829	0.843	<u>0.750</u>	**0.888**
(Power)	700/700	0.990	0.995	<u>0.984</u>	**0.985**
	2,000/700	**0.999**	**0.999**	<u>0.998</u>	<u>0.998</u>
	2,000/2,000	**<u>1.000</u>**	**<u>1.000</u>**	**<u>1.000</u>**	**<u>1.000</u>**

Note. The most powerful strategy's power rate is bolded while the least powerful strategy's power rate is underlined. R/F = reference/focal.

[a] Nonrobust Type I error rates that are outside the 0.025 to 0.075 range.

Discussion

The perspective of this study is that conditional DIF assessments are useful for evaluating an item's DIF at a more detailed level than summary significance tests and effect sizes. This more detailed level can be important when summary DIF assessments oversummarize an item's extent of DIF or summarize DIF when the summary is not of direct interest. The focus of the study was on evaluating the accuracy of four estimation strategies with respect to their conditional DIF estimates, with a secondary focus on these estimation strategies' accuracies in estimating a common DIF effect size and their statistical significance tests.

The overall results suggested that the logistic regression and loglinear models' strategies were the most and second most recommended of the four evaluated DIF estimation strategies. The logistic regression estimation strategy was especially useful for estimating conditional DIF in small sample sizes and for a powerful statistical significance test of overall DIF. The loglinear models' estimation strategy could approximate the conditional DIF in the population better than the logistic regression estimation strategy when the population's conditional DIF was complex, however, it required large sample sizes for its flexibility to outweigh its relatively large standard errors and its reduced statistical power. The loglinear models' estimation strategy offers a wider range of parameterizations than logistic regression (Appendix C), where increasing the number of parameters in the loglinear models from what was used in this study can approximate data even more closely, while decreasing the number of parameters can reduce standard errors and perhaps increase statistical power. The decision process for selecting appropriate parameterizations in the loglinear models' strategy can be very extensive (e.g., Hanson & Feinstein, 1995). The raw data strategy produced conditional DIF estimates that were relatively unbiased with respect to the population's conditional DIF, but also had high levels of variability that cause conditional DIF assessments to elude interpretation for all but the largest sample sizes. The raw data, logistic regression, and loglinear models estimated the standardized E-Dif measure of overall DIF with almost identical levels of accuracy.

The performance of kernel smoothing made it the least desirable of the four considered DIF estimation strategies. It produced the most biased conditional DIF estimates of the four considered estimation strategies, had a significance test with an inflated Type I error rate, and was the only strategy with bias levels large enough to reduce the accuracy of the overall standardized E-Dif measure to levels of practical concern. The source of kernel smoothing's

inaccuracy is that it smoothes the $E(Y_m)$'s separately for the reference and focal groups, meaning that the groups' smoothing parameters and extent to which each of the *M* levels of $E(Y_m)$ are weighted in its weighted averaging process are a direct function of the groups' overall and conditional sample sizes (Appendix D). When the groups differ in their overall ability, the $E(Y_m)$'s that are closely fit and strongly smoothed are different across the groups, creating inaccuracy in the conditional DIF estimates that inflates bias and Type I error rates. The effect of reference and focal group differences on the accuracy of kernel smoothing can be observed in the higher Type I error rates, conditional biases, and overall biases of the science items than the history items (Tables 3, 6, and 9), as the science items' data exhibited larger reference and focal differences than the history items' data (Table 2).

Some follow-up efforts were made to try to improve the application of kernel smoothing in DIF assessments; one involved smoothing the raw conditional DIF estimates and another used a single weighting function to smooth both the reference and focal $E(Y_m)$'s. These efforts did not improve kernel smoothing beyond the version assessed in this study and even created additional inaccuracies which would be difficult to address (such as how to deal with one group's missing data at an X_m score).

Future Directions

Some issues not considered in this study could be the basis of future studies. The current study compared the DIF strategies under simple conditions where all of the items making up the X_m score could be assumed to be non-DIF items. Future studies could evaluate the performance of the DIF estimation strategies when used with all items on the test making up X_m (including Y) or when used with a data-based purification approach where all of the items on the test are evaluated for DIF and then the DIF items are excluded from X_m when evaluating Y. Wider ranges of reference and focal group sample sizes could also be considered.

An important extension of this investigation is to the evaluation of conditional DIF in polytomous items. The features of polytomous items would likely accentuate the differences between the loglinear models and logistic regression strategies. The loglinear models' strategy would require several parameters to model the frequency distributions of each possible score on the studied item, probably reducing its statistical power and making model convergence less

likely for small and moderate sample sizes. The unconstrained cumulative logits version of logistic regression has been demonstrated to have an accurate Type I error and high power as an overall significance test (Kristjansson et al., 2005), implying that its conditional DIF estimates would be most recommended.

One DIF situation that could form an important follow-up study is a nonuniform DIF situation where the conditional DIF crosses to such an extent that the overall standardized E-Dif is close to zero. It may not be likely to find such a situation in practice, and even if found, this situation might be more likely explained by sampling variability than by substantive explanation. However, an extreme crossing DIF situation could be an important basis for studying the differences among the four DIF estimation strategies' significance tests and null hypotheses. Specifically the logistic regression and loglinear models' strategies explicitly incorporate nonuniform DIF into their test statistics, perhaps making them more likely to detect crossing DIF than the raw data and kernel smoothed standardized strategies that focus on testing the standardized E-Dif.

Some readers might be more interested in assessing DIF that is defined in terms of an expected true score matching variable (Shealy & Stout, 1993) than in terms of an observed score matching variable (1). While the SIBTEST approach to DIF is different from that considered in this study, the logistic regression and loglinear models' estimation strategies have potential to work within and improve the SIBTEST procedure. Moses and Miao (2007) have shown that the use of loglinear models for estimating conditional DIF rather than raw data provides stability that allows the SIBTEST regression correction to work more closely to how it is intended to work. The use of loglinear models, and potentially logistic regression models, also avoids and possibly improves on the use of data exclusion strategies that have been advocated for the SIBTEST procedure (Shealy & Stout).

A final discussion point is how the DIF criteria used in this study affected how well the DIF strategies performed. As stated throughout this study's Method section, the DIF criteria chosen in this study were the DIF values computed from large populations of raw test data. Reviewers of this study have expressed concerns that this study's populations of raw test data may have advantaged some of the considered strategies (i.e., raw data, loglinear models) over others (i.e., logistic regression). These reviewer concerns can be informed by an awareness that comparative studies of DIF methods always require a choice of how the DIF criteria and

populations are defined. In prior DIF studies, DIF methods have been compared based on criteria and populations ranging from actual test data (Dorans & Holland, 1993; Hanson & Feinstein, 1995; Lyu et al., 1995; Miller & Spray, 1993; Moses & Miao, 2007; Puhan et al., 2007) to data that have been simulated with degrees of nonuniform DIF and with presumed relationships between observed scores and latent variables (Douglas et al., 1996; Kristjansson et al., 2005; Roussos & Stout, 1996; Shealy & Stout, 1993; Swaminathan & Rogers, 1990).

Because a choice is required for how criteria and populations are defined in DIF studies, justifications of these choices can be useful for interpreting DIF studies, their motivations, and their results. The justifications for the current study's use of DIF values computed from large samples of raw test data as DIF criteria are that 1) large sample DIF criteria are realistic and therefore relevant for practice (as stated in this study's Method section), and 2) all four of the considered DIF strategies have been recommended and used to estimate DIF in actual test data but have not been extensively compared (as stated in this study's introduction). Additional investigations could be undertaken to address concerns that one or more of this study's considered strategies was disadvantaged by this study's use of realistic DIF criteria. The additional investigations could focus on comparing DIF estimation strategies with respect to artificial criteria that directly cater to strategies such as logistic regression (i.e., logistic item response functions rather than observed item response functions).

References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144–152.

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., & 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*(4), 355–368.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–165). Hillsdale, NJ: Lawrence Erlbaum Associates.

Douglas, J. A., Stout, W., & DiBello, L. V. (1996). A kernel-smoothed version of SIBTEST with applications to local DIF inference and function estimation. *Journal of Educational and Behavioral Statistics, 21*(4), 333–363.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315–332.

Gierl, M. J., & Bolt, D. M. (2001). Illustrating the use of nonparametric regression to assess differential item and bundle functioning among multiple groups. *International Journal of Testing, 1*(3&4), 249–270.

Hanson, B. A., & Feinstein, Z. S. (1995, April). *A polynomial loglinear model for assessing differential item functioning.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*, 133–183.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49*(2), 223–245.

Keselman, H. J., Wilcox, R. R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods, 1*(2), 288–309.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935–953.

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

Lyu, C. F., Dorans, N. J., & Ramsay, J. O. (1995). *Smoothed standardization assessment of testlet level DIF on a math free-response item type* (ETS Research Rep. No. RR-95-38). Princeton, NJ: ETS.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105–118.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*(2), 107–122.

Moses, T., & Miao, J. (2007, April). *An incisive look at a suspect regression correction.* Paper presented at the annual meeting of the National Council for Measurement in Education, Chicago, IL.

Puhan, G., Moses, T., Yu., L., & Dorans, N. (2007, April). *Small-sample DIF estimation using log-linear smoothing: A SIBTEST application.* Paper presented at the annual meeting of the National Council for Measurement in Education, Chicago, IL.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611–630.

Ramsay, J. O. (2000). TESTGRAF: A program for the graphical analysis of multiple-choice test and questionnaire data [Computer software and manual]. Retrieved from http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html

Rao, C. R. (1966). *Linear statistical inference and its applications*. New York, NY: Wiley.

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

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

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9–30). Baltimore, MD: Johns Hopkins University Press.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361370.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type item scores*. Retrieved November 1, 2009, from Ottawa, Canada, Directorate of Human Resources Research and Evaluation, Department of National Defense Web site: http://www.educ.ubc.ca/faculty/zumbo/DIF/index.

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*(2), 225–247.

List of Appendices

Appendix A

Differential Item Functioning (DIF) Estimates Using Raw Data

The reference and focal expected scores of (1) and (2) can be estimated as the sample means from the raw data

$$E(Y_{Fm}) = \frac{\sum_{i \text{ in } F \text{ and } m} Y_{iFm}}{n_{Fm}} \quad \text{and} \quad E(Y_{Rm}) = \frac{\sum_{i \text{ in } R \text{ and } m} Y_{iRm}}{n_{Rm}}, \tag{A1}$$

with estimated variances from the raw data,

$$Var(E(Y_{Fm})) = \frac{\sum_{i \text{ in } F \text{ and } m} (Y_{iFm} - E(Y_{Fm}))^2}{n_{Fm}^2} \quad \text{and} \quad Var(E(Y_{Rm})) = \frac{\sum_{i \text{ in } R \text{ and } m} (Y_{iRm} - E(Y_{Rm}))^2}{n_{Rm}^2}. \tag{A2}$$

The standard error of (2) can be estimated as,

$$\sqrt{\sum_m \left(\frac{n_{Fm}}{\sum_m n_{Fm}} \right)^2 \left(Var(E(Y_{Fm})) + Var(E(Y_{Rm})) \right)}, \tag{A3}$$

(Dorans & Holland, 1993, p. 50). The division of (A3) into (2) has been promoted as a *z*-test of DIF in (2) (e.g., the *z*-test of the SIBTEST version of the standardized E-Dif is described in Shealy & Stout, 1993, p. 169).

Appendix B

Differential Item Functioning (DIF) Estimates Using Logistic Regression

The application of logistic regression procedures to DIF assessment (French & Miller, 1996; Jodoin & Gierl, 2001; Kristjansson et al., 2005; Swaminathan & Rogers, 1990) involves predicting the probability of a correct response (= 1) on dichotomously-scored *Y* based on total score, X_m, and group membership. Logistic models of the separate reference and focal groups' predicted *Y*'s can be estimated and directly used in (1) and (2) as the $E(Y)$'s,

$$P(Y_{Rm}=1|X_m)=\frac{1}{1+e^{-\beta_{0R}-\beta_{1R}X_m}}=\frac{1}{1+e^{-\boldsymbol{\beta}_{\mathbf{R}}^{\mathbf{t}}\mathbf{D_m}}} \text{ and}$$

$$P(Y_{Fm}=1|X_m)=\frac{1}{1+e^{-\beta_{0F}-\beta_{1F}X_m}}=\frac{1}{1+e^{-\boldsymbol{\beta}_{\mathbf{F}}^{\mathbf{t}}\mathbf{D_m}}}, \quad \text{(B1)}$$

The β terms in the models are estimated by maximum likelihood. The rightmost expressions of (B1) are matrix expressions helpful for additional derivations, where $\boldsymbol{\beta}^{\mathbf{t}}$ is the transposed row vector of β_0 and β_1 terms, (β_0,β_1), and $\mathbf{D_m}$ is the *m*th 2-by-1 design matrix containing 1 and X_m, $\begin{pmatrix}1\\X_m\end{pmatrix}$.

Estimates of the variances of the $E(Y)$'s for (3) can be computed from (B1) based on differentiating the functions and applying the delta method. When $P(Y_{Rm}=1|X_m)$ is used as $E(Y_{Rm})$,

$$\begin{aligned}Var(E(Y_{Rm}))&=\left(\frac{\partial P(Y=1|X_m)}{\partial\boldsymbol{\beta}_{\mathbf{R}}}\right)^{\mathbf{t}}Var(\boldsymbol{\beta}_{\mathbf{R}})\frac{\partial P(Y=1|X_m)}{\partial\boldsymbol{\beta}_{\mathbf{R}}}\\&=\left(\frac{e^{-\boldsymbol{\beta}_{\mathbf{R}}^{\mathbf{t}}\mathbf{D_m}}}{\left(1+e^{-\boldsymbol{\beta}_{\mathbf{R}}^{\mathbf{t}}\mathbf{D_m}}\right)^2}\mathbf{D_m}\right)^{\mathbf{t}}Var(\boldsymbol{\beta}_{\mathbf{R}})\left(\frac{e^{-\boldsymbol{\beta}_{\mathbf{R}}^{\mathbf{t}}\mathbf{D_m}}}{\left(1+e^{-\boldsymbol{\beta}_{\mathbf{R}}^{\mathbf{t}}\mathbf{D_m}}\right)^2}\mathbf{D_m}\right),\end{aligned} \quad \text{(B2)}$$

where the 2-by-2 variance-covariance matrix $Var(\boldsymbol{\beta}_{\mathbf{R}})$ is the negative inverse of the second derivatives of the $P(Y_{Rm}=1|X_m)$ model's loglikelihood function with respect to the model's parameters, $\boldsymbol{\beta}_{\mathbf{R}}$, when the maximum likelihood algorithm converges (Rao, 1966). The estimation of $Var(E(Y_{Fm}))$ is similar.

The logistic regression's overall significance test is based on modeling the probability of a correct response (=1) on *Y* using both the reference and focal data in overall models with total score X_m, a dichotomously-coded group membership variable, G_m, and the interaction of group membership and X_m, $X_m G_m$. One model allows for DIF by expressing the separate reference and focal models in (B1) in an overall model,

$$P(Y_m=1|X_m,G_m,X_mG_m)=\frac{1}{1+e^{-\beta_0-\beta_1X_m-\beta_2G_m-\beta_3X_mG_m}}. \qquad \text{(B3)}$$

Another model constrains *Y*'s DIF to be zero in the reference and focal data,

$$P(Y_m=1|X_m)=\frac{1}{1+e^{-\beta_0-\beta_1X_m}}. \qquad \text{(B4)}$$

Model (B3) is a nonuniform DIF model that models *Y* based partly on constant reference and focal group differences (β_2G_m) across X_m and partly on reference and focal group differences that are allowed to vary with X_m ($\beta_3X_mG_m$). The logistic framework provides its own significance test for nonuniform DIF using the likelihood ratio test comparing models (B3) and (B4),

$$\chi^2=-2(\ln L(M_{B4})-\ln L(M_{B3})), \qquad \text{(B5)}$$

where $\ln L(M_{B4})$ is the maximized loglikelihood for model (B4),

$$\ln L(M_{B4})=\sum_m\left(n_{R+F,m,1}\ln P(Y_m=1|X_m)+n_{R+F,m,0}\ln P(Y_m=0|X_m)\right), \qquad \text{(B6)}$$

and $n_{R+F,m,1}$ and $n_{R+F,m,0}$ are the numbers of reference and focal examinees at score X_m that obtain 1 and 0 on *Y*, respectively. $\ln L(M_{B3})$ is defined similarly. The statistic in (B5) is chi-square distributed with degrees of freedom equal to the difference in the degrees of freedom for models (B3) and (B4), or 2.

Appendix C

Differential Item Functioning (DIF) Estimates Using Loglinear Models

Loglinear models are used to separately estimate the frequency distributions of the DIF matching variable X_m for each response category of *Y*. For the reference examinees who get *Y* correct (=1), the frequency distribution of X_m can be modeled as,

$$\ln(s_{RmY=1}) = \beta_0 + \sum_{v=1}^{V} \beta_v X_m^v \quad , \qquad \text{(C1)}$$

where $s_{RmY=1}$ is the expected (not actual) frequency of reference examinees who get *Y* correct and obtain score X_m and the β terms are estimated using maximum likelihood (Holland & Thayer, 2000). The *V* is chosen by the modeler and must be less than the total number of scores on X_m, *M*. The maximum likelihood estimation of model (C1) produces a smoothed frequency distribution $s_{RmY=1}$, where the first *V* moments (mean, variance, skewness, etc.) match those of the observed frequency distribution, $n_{RmY=1}$. *V* is set at 4 for all models and conditions of this study.

The $E(Y_m)$'s are computed based on the separate modeling of four X_m frequency distributions, $s_{RmY=1}$, $s_{RmY=0}$, $s_{FmY=1}$ and $s_{FmY=0}$, with loglinear models such as (C1),

$$E(Y_{Fm}) = \frac{s_{FmY=1}}{s_{FmY=1} + s_{FmY=0}} \quad \text{and} \quad E(Y_{Rm}) = \frac{s_{RmY=1}}{s_{RmY=1} + s_{RmY=0}} \, . \qquad \text{(C2)}$$

The $E(Y_m)$'s from (C2) are used in (1) and (2).

Estimates of the variances of the $E(Y)$'s for (3) can be computed from (C2) based on the delta method. For $E(Y_{Rm})$,

$$
\begin{aligned}
&Var(E(Y_{Rm})) \\
&= \frac{\partial E(Y_{Rm})}{\partial s_{RmY=1}} Var(s_{RmY=1}) \frac{\partial E(Y_{Rm})}{\partial s_{RmY=1}} + \frac{\partial E(Y_{Rm})}{\partial s_{RmY=0}} Var(s_{RmY=0}) \frac{\partial E(Y_{Rm})}{\partial s_{RmY=0}} \\
&= \left(\frac{s_{RmY=0}}{\left(s_{RmY=1} + s_{RmY=0}\right)^2} \right)^2 Var(s_{RmY=1}) + \left(\frac{-s_{RmY=1}}{\left(s_{RmY=1} + s_{RmY=0}\right)^2} \right)^2 Var(s_{RmY=0})
\end{aligned}
\tag{C3}
$$

where $Var(s_{RmY=1})$ is obtained from the $s_{RmY=1}$ model's results and is the *m*th diagonal entry of $= \left(\frac{\partial \mathbf{s}_{\mathbf{RY=1}}}{\partial \boldsymbol{\beta}_{\mathbf{RY=1}}} \right) \mathbf{Var}(\boldsymbol{\beta}_{\mathbf{RY=1}}) \left(\frac{\partial \mathbf{s}_{\mathbf{RY=1}}}{\partial \boldsymbol{\beta}_{\mathbf{RY=1}}} \right)^{\mathbf{t}} = \left(\boldsymbol{\Sigma}_{\mathbf{s}_{\mathbf{RY=1}}} \mathbf{D}_{\mathbf{RY=1}} \right) \mathbf{Var}(\boldsymbol{\beta}_{\mathbf{RY=1}}) \left(\mathbf{D}_{\mathbf{RY=1}}^{\mathbf{t}} \boldsymbol{\Sigma}_{\mathbf{s}_{\mathbf{RY=1}}} \right)$, where $\boldsymbol{\Sigma}_{\mathbf{s}_{\mathbf{RY=1}}} = \mathbf{DIAG}_{\mathbf{s}_{\mathbf{RY=1}}} - N_R^{-1} \mathbf{s}_{\mathbf{RY=1}} \mathbf{s}_{\mathbf{RY=1}}^{\mathbf{t}}$, $\mathbf{DIAG}_{\mathbf{s}_{\mathbf{RY=1}}}$ is the diagonalized matrix of $\mathbf{s}_{\mathbf{RY=1}}$, $\mathbf{D}_{\mathbf{RY=1}}$ is an M+1-by-V design matrix containing all of the $s_{RmY=1}$ model's X_m^v terms, and $\mathbf{Var}(\boldsymbol{\beta}_{\mathbf{RY=1}})$ is the negative inverse of the second derivatives of the $s_{RmY=1}$ model's loglikelihood function with respect to the model's parameters, $\boldsymbol{\beta}_{\mathbf{RY=1}}$, when the maximum likelihood algorithm converges (Holland & Thayer, 2000). The estimation of $Var(s_{RmY=0})$ is similar to that of $Var(s_{RmY=1})$. The estimation of $Var(E(Y_{Fm}))$ is similar to that of $Var(E(Y_{Rm}))$.

Overall models of the X_m frequency distributions of the focal and reference data for the two possible scores on *Y* can be fit to create statistical significance tests of *Y*'s DIF. Let G_m be a dichotomously coded indicator of focal or reference group membership and let Y_m indicate the obtained score on *Y*, where both levels of G_m and Y_m appear for all levels of X_m. Two models considered in this study are a nonuniform DIF model that combines all of the independently modeled $s_{RmY=1}$, $s_{RmY=0}$, $s_{FmY=1}$ and $s_{FmY=0}$ distributions of form (C1) into an overall model,

$$
\begin{aligned}
&\ln(s_{GmY}) = \\
&\beta_0 + \sum_{v=1}^{V} \beta_{Xv} X_m^v + \beta_Y Y_m + \beta_G G_m + \sum_{v=1}^{V} \beta_{X,Y,v} X_m^v Y_m + \sum_{v=1}^{V} \beta_{X,G,v} X_m^v G_m + \beta_{Y,G} Y_m G_m + \sum_{v=1}^{V} \beta_{X,Y,G,v} X_m^v Y_m G_m
\end{aligned}
\tag{C4}
$$

and a non-DIF model,

$$\ln(s_{GmY}) = \beta_0 + \sum_{v=1}^{V} \beta_{Xv} X_m^v + \beta_Y Y_m + \beta_G G_m + \sum_{v=1}^{V} \beta_{X,Y,v} X_m^v Y_m + \sum_{v=1}^{V} \beta_{X,G,v} X_m^v G_m. \quad \text{(C5)}$$

Model (C5) does not contain (C4)'s terms that allow for uniform DIF that is constant across the X_m categories, $\beta_{Y,G} Y_m G_m$, and nonuniform DIF that allows DIF to vary across the X_m categories, $\sum_{v=1}^{V} \beta_{X,Y,G,v} X_m^v Y_m G_m$. There are many variations on these two models for assessing DIF, and some of the implications of using other models are described in the Discussion section.

A significance test of DIF can be computed by comparing the loglikelihoods of models (C4) and (C5),

$$\chi^2 = -2(\ln L(M_{C5}) - \ln L(M_{C4})), \quad \text{(C6)}$$

where $\ln L(M_{C5})$ is the maximized loglikelihood for model (C5),

$$\ln L(M_{C5}) = \sum_Y \sum_G \sum_m n_{GmY} \ln\left(\frac{s_{GmY}}{\sum_Y \sum_G \sum_m s_{GmY}}\right). \quad \text{(C7)}$$

The statistic in (C6) is chi-square distributed with degrees of freedom equal to the difference in the degrees of freedom for models (C5) and (C4), or $V + 1$.

Appendix D

Differential Item Functioning (DIF) Estimates Using Kernel Smoothing

Kernel smoothing computes kernel-smoothed $E(Y_m)$'s as moving and weighted averages of the raw $E(Y_m)$'s estimated in (A1). These kernel smoothed expected scores, $KSE(Y_{Rm})$ and $KSE(Y_{Fm})$, can be used in (1) and (2),

$$KSE(Y_{Rm}) = \mathbf{w_{Rm}E(Y_R)} \text{ and } KSE(Y_{Fm}) = \mathbf{w_{Fm}E(Y_F)}, \quad \text{(D1)}$$

where the $\mathbf{E(Y_R)}$ and $\mathbf{E(Y_F)}$ are *M* row vectors containing each of the raw $E(Y_m)$'s, and $\mathbf{w_{Rm}}$ and $\mathbf{w_{Fm}}$ are 1-by-*M* matrices each containing *l* = 1 to *M* kernel weights, $w_{Rm,l}$ and $w_{Fm,l}$. The kernel weights considered here are Gaussian weights,

$$w_{Rm,l} = \frac{e\left[\frac{-1}{2h}\left(\frac{X_l - X_m}{\sigma_{XR}}\right)^2\right] n_{Rl}}{\sum_l e\left[\frac{-1}{2h}\left(\frac{X_l - X_m}{\sigma_{XR}}\right)^2\right] n_{Rl}}, \quad \text{(D2)}$$

which are one type of kernel smoothing weights that include and are understood to perform similarly to quadratic, uniform, logistic weights (Douglas et al., 1996; Ramsay, 1991). In (D2), n_{Rl} is the reference group's sample size at X_l, σ_{XR} is the reference group's standard deviation on *X*, and *h* is a kernel bandwidth parameter that determines the extent of smoothing done to the $E(Y_m)$'s in computing the $KSE(Y_m)$'s. Suggestions of default *h* values are typically based on total sample size (e.g., Douglas et al., 1996; Ramsay, 1991, p. 618). In this study h is set at $1.1N^{-.2}$, where *N* is the reference group's total sample size. The kernel weights for the focal group, $w_{Fm,l}$, are computed similarly to $w_{Rm,l}$ by using the focal group's conditional and overall sample sizes and the focal group's standard deviation of *X*. The kernel weights given in (D2) are how kernel smoothing is done at ETS to assess item response functions without the use of parametric models and also to assess conditional DIF.

The variances of (D1) that can be used in (3) can be computed using the raw conditional variances estimated in (A2) and the kernel weighting functions in (D2),

$$Var(KSE(Y_{Rm})) = \mathbf{w_{Rm}Var(E(Y_R))w_{Rm}^t} \text{ and } Var(KSE(Y_{Fm})) = \mathbf{w_{Fm}Var(E(Y_F))w_{Fm}^t}. \quad \text{(D3)}$$

In (D3), the $\mathbf{Var(E(Y_F))}$ and $\mathbf{Var(E(Y_R))}$ are *M*-by-*M* matrices containing the *M* raw conditional variances in the diagonal cells and zeros in the other cells.

The estimate of the standard error for an overall kernel-smoothed standardized E-Dif statistic can be obtained by expressing the kernel-smoothed standardized E-Dif statistic based on using the kernel-smoothed terms in (D1) in (2),

$$KSStnd = \left(\frac{1}{N_F}\right)\mathbf{n_F^t}\left(\mathbf{w_F E(Y_F) - w_R E(Y_R)}\right), \quad \text{(D4)}$$

and then applying the delta method,

$$Var(KSStnd) = \left(\frac{1}{N_F}\right)\mathbf{n_F^t w_F Var(E(Y_F))w_F^t n_F}\left(\frac{1}{N_F}\right) + \left(\frac{1}{N_F}\right)\mathbf{n_F^t w_R Var(E(Y_R))w_R^t n_F}\left(\frac{1}{N_F}\right). \quad \text{(D5)}$$

In (D4) and (D5), N_F is the total sample size of the focal group, $\mathbf{n_F^t}$ is the transposed *M*-by-1 vector of the focal group's observed frequencies at all *M* score levels of X_m, and $\mathbf{w_F}$ and $\mathbf{w_R}$ are *M*-by-*M* matrices containing all *M* 1-by-*M* $\mathbf{w_{Rm}}$ and $\mathbf{w_{Fm}}$ matrices stacked from *m* = 1 to *M*. This study evaluates the accuracy of a z-test of (D4) based on dividing it by the square root of (D5).