CollegeBoard
connect to college success™

Research Report
No. 2005-10

Using DIF Dissection Method to Assess Effects of Item Deletion

Yanling Zhang, Neil J. Dorans, and
Joy L. Matthews-López

www.collegeboard.com

College Board Research Report No. 2005-10
ETS RR-05-23

Using DIF Dissection Method to Assess Effects of Item Deletion

Yanling Zhang, Neil J. Dorans, and Joy L. Matthews-López

The College Board, New York, 2005

Yanling Zhang is a measurement statistician in the Center of Statistical Analysis at Educational Testing Service.

Neil J. Dorans is a distinguished presidential appointee at Educational Testing Service.

Joy L. Matthews-López is director of research in the Centers for Osteopathic Research and Education at Ohio University. She previously held the position of associate measurement statistician in the Center of Statistical Analysis at Educational Testing Service.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board: Connecting Students to College Success

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 5,000 schools, colleges, universities, and other educational organizations. Each year, the College Board serves seven million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #050481690) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Acknowledgments

This research was funded by the College Board. The authors wish to express their thanks to the College Board for allowing them to use SAT Reasoning Test™ verbal section data for this research. The opinions expressed herein are those of the authors and should not be interpreted as ETS or College Board policy statements. An earlier version of this paper was presented at the 2003 annual meeting of the National Council on Measurement in Education in Chicago, Illinois. The authors also thank Keith Boughton, Dan Eignor, Jennifer Kobrin, Glenn Milewski, and Saba Rizavi for their helpful comments.

Contents

Tables

Figures

Abstract

Statistical procedures for detecting differential item functioning (DIF) are often used as an initial step to screen items for construct irrelevant variance. This research applies a DIF dissection method and a two-way classification scheme to SAT Reasoning Test™ verbal section data and explores the effects of deleting sizable DIF items on reported scores after re-equating. The DIF dissection approach and the two-way classification scheme may yield new and detailed insight into item functioning at the subgroup level. Two hypotheses are studied: (1) whether or not the deletion of a sizeable DIF item that is the most disadvantageous to a particular subgroup will affect the scores for that subgroup the most; and (2) whether or not the effects of item deletion on scores can be predicted by the standardization method. Both hypotheses are supported by the results of this research.

Keywords: SAT®, DIF, standardization method, DIF dissection approach, two-way DIF classification

Background

Standardized achievement tests often have high stakes attached to their use. Statistical procedures for detecting differential item functioning (DIF) are frequently used as an initial step to screen items for construct irrelevant variance. First of all, it is necessary to distinguish between DIF and impact. DIF indicates a difference in item performance between two comparable groups of examinees, i.e., groups that are matched with respect to the construct being measured by the test. On the other hand, impact refers to a difference in item and test performance between two intact groups. Standard DIF detection procedures focus on only one categorical variable at an aggregated group level, such as gender or ethnicity/race. To date, DIF studies in the arena of standardized achievement testing have investigated gender separately from ethnicity/race (e.g., Carlton and Harris, 1992; Doolittle and Cleary, 1987; O'Neil and McPeek, 1993; Scheuneman and Grima, 1997; and Schmitt and Dorans, 1990).

Hu and Dorans (1989) used data from the SAT Reasoning Test verbal section to examine the effect of deleting both minimal and sizable DIF items on equating functions and subsequent reported scores. The hypothesis they tested was whether the deletion of minimal and/or sizable DIF items resulted in different scaled scores after IRT true score re-equating and Tucker re-equating (Kolen and Brennan, 2005). The results of that study indicated that the act of deleting the item itself had a noticeable effect on scaled scores. The effect size of a DIF item had a less prominent effect on the scaled scores.

Dorans and Holland (1993) pointed out that in traditional one-way DIF analysis, deleting items due to DIF can have unintended consequences on the focal group. DIF analysis performed on gender and on ethnicity/race alone ignores the potential interactions between the two main effects. Additionally, Dorans and Holland suggested applying a "melting pot" DIF method wherein the total group would function as the reference group and each gender-by-ethnic subgroup would serve sequentially as a focal group.

Zhang (2001) argued that DIF analysis with a traditional one-way approach does not serve the purpose of illuminating actual gender and ethnic/racial performance differences. A two-way DIF classification scheme was proposed in which each item was examined for DIF effect at the subgroup level, i.e., gender DIF within ethnicity/race and ethnicity/race DIF within gender. The results of that study identified several gender and ethnic/racial DIF items which were previously undetected in a total analysis and yet were flagged when two-way procedures were applied.

Research Questions

Building on previous work by Dorans and Holland (1993), Hu and Dorans (1989), and Zhang (2001), this research applies a "melting pot" DIF approach and a two-way classification scheme to SAT Reasoning Test verbal section data. Subsequently, the effect of deleting sizable DIF items on reported scores after equipercentile re-equating is examined. As mentioned earlier, this melting pot DIF approach utilizes nontraditional reference and focal-group formations. For purposes of this research, this approach will be referred to as "DIF dissection." In DIF dissection, each subgroup will act as an independent focal group while the total group will function as the reference group. In essence, the total group is dissected into a set of complementary focal groups. Since we were investigating the effect of DIF items on subgroups, using the total group that contains the focal groups gives a common benchmark for DIF comparisons. The DIF dissection analysis takes the interactions between gender and ethnicity/race into consideration. Regular DIF analysis done in SATs ignores these interactions. The new ETS "Fairness Guidelines" on ethnocentrism argue against using a group like males or whites as references for other groups. So using the total group as the reference is less ethnocentric.

There are two goals of this research: (1) to examine items for DIF using the DIF dissection method and the two-way classification scheme described above within the standardization DIF detection procedure, and (2) to assess the effect, if any, of deleting items with sizable DIF statistics on the reported scores for all groups after re-equating the shortened tests.

The hypotheses to be tested are the following:

1. The deletion of a DIF item that is the most disadvantageous to a particular subgroup will affect the scores for that subgroup the most;
2. The effects of item deletion on scores can be predicted by measures of DIF in the item's score metric that can be produced by the standardization method.

Method

Data Source

Data were obtained from a spring 2001 administration of the SAT. All test editions consisted of 78 five-option multiple-choice verbal items. In addition to these operational items, each test contained a 30-minute, nonoperational section that was used for equating purposes as well as for pretesting new items. This research is limited to the use of multiple-choice verbal pretest sections that consist of 35 items. Instructions to test-takers directed them to choose the best of the five provided options for each item.

For this research, examinees were classified by both gender and ethnicity/race. Following the subgroup classification scheme used by Dorans and Holland (2000), all examinees that indicated their gender but not their ethnicity/race in a group were labeled as "All Others." In addition, Native Americans were also placed in "All Others" since this particular sample was too small for subgroup-level analyses.

A total of 10 subgroups were formed: African American Females, African American Males, Asian Females, Asian Males, Hispanic Females, Hispanic Males, White Females, White Males, All Other Females, and All Other Males (see Table 1 below). For purposes of DIF analyses, the reference group was defined to be the total group; the focal groups were formed according to each of the 10 subgroups (see Table 1).

The Mantel-Haenszel DIF statistic (Holland and Thayer, 1988) and the ETS Mantel-Haenszel delta-difference criteria (Zieky, 1993) were used to flag and classify DIF items in the pretest forms. The ETS criteria are explained later in this section. The authors started the project by reviewing the DIF summary statistics for all verbal and mathematics pretest forms from a single administration of the SAT. These summary statistics were reviewed in terms of the number of items with sizable DIF (C-level) and the degree of DIF effects. Specific verbal sections were chosen for further screening if items with more sizable DIF were flagged. Of the different pretests, only one was retained for this research because it had six C-level (sizable) DIF items.

Table 1

Composition of Reference Group and Focal Groups

Reference Group	*Focal Groups*	
Total Group	African American Female	African American Male
	Asian Female	Asian Male
	Hispanic Female	Hispanic Male
	White Female	White Male
	All Other Female	All Other Male

Three out of the six C-level items in the pretest form were selected to be analyzed in detail using the standardization DIF procedure. The standardization method (Dorans and Kulick, 1986) was chosen for this work because it is easily adapted to a formula-scored test as well as to the scenario of multigroup analyses. It also lends itself well to the prediction of effects of item deletion on subgroup performance. Each of the three sizable DIF items in this particular pretest was removed from the responses of the affected group as well as from all other groups. Equipercentile equating was performed to re-equate the shortened pretest test to the 78-item operational test after each item deletion. The full pretest was also equated to the 78-item operational test. Smoothing was not used because the sample size in this research was sufficiently large. Reported score distribution and score changes of each ethnic and gender group were then examined after the systematic deletion of each item that had sizable DIF.

Formula-Scoring Procedures

The scoring procedure for the SAT utilizes a formula-scoring (FS) procedure and is defined as follows:

$$FS = Rights * 1 + (\text{Omits and Not Reached}) * 0 + (\frac{-1}{k-1}) * Wrongs,$$

where k equals the number of options for each multiple-choice item. As can be seen, omitted and not-reached items (NR) are treated differently than incorrect responses. Whereas points are neither awarded nor deducted for omitted and not-reached items, incorrect responses to the multiple-choice items result in a loss of a fraction of a point. For five-choice items, each incorrect response results in a 0.25 deduction from the total FS score.

DIF Detection Procedure—The Standardization Method

The standardization method (STD) for DIF detection (Dorans and Kulick, 1986; Dorans and Schmitt, 1993) was used in this study. As stated by Dorans and Holland (1993), the standardization method is readily adapted to formula-scored items, such as those used on the SAT verbal section. Since the SAT is a formula-scored test, DIF given by the standardization method indices, *STD FS-DIF*, was used for evaluation in this study. Using a formula-scored DIF procedure for a formula-scored test provides consistent conditions under which the item was analyzed. *STD FS-DIF* incorporates a formula scoring algorithm and assigns zero weight to omitted and not-reached items, and [$\frac{-1}{k-1}$] to incorrect responses, where k is the number of answer options. The *STD FS-DIF* index ranges between –1.25 to +1.25, inclusive in this case where $k = 5$.

One of the main principles underlying the standardization approach is to use all available appropriate data to estimate the conditional item performance of each group at each level of the matching variable. An item exhibits DIF when the expected performance on an item differs for matched examinees from different groups. Expected performance can be operationalized by nonparametric item-test regressions. Differences in empirical item-test regressions are indicative of DIF.

The standardization definition of DIF at the individual score level, m, is given by $D_m = FS_{fm} - FS_{rm}$, where FS_{rm} are item-test regressions at the score level m for the focal group and the reference group respectively. For formula-scored items, STD has a DIF index defined by the standardized formula score-difference (*STD FS-DIF*), given by

$$STD\ FS\text{-}DIF = \frac{\sum_{m=1}^{M}\left[N_{fm}\left(FS_{fm} - FS_{rm}\right)\right]}{\sum_{m=1}^{M} N_{fm}}$$

where, $\dfrac{N_{fm}}{\sum_{m=1}^{M} N_{fm}}$

is the weighting factor at score-level m. Score-level m is supplied by the standardization group (which is each individual focal group in this case) to weight differences in item performance between the focal group, FS_{fm}, and the reference group, FS_{rm}.

The ETS DIF Classification Scheme

Zieky (1993) described the ETS DIF classification scheme for use in test development. To paraphrase Zieky, for a certain combination of reference and focal groups, all the items can be categorized into one of three groups:

1) Category A, which refers to items either for which the magnitude of delta$_{\text{MH}}$ values is < 1 delta unit in absolute value or for which delta$_{\text{MH}}$ is not statistically significantly different from 0;
2) Category C, which refers to items with delta$_{\text{MH}}$ > 1.5 delta units in absolute value and are statistically significantly > 1.0 in absolute value; and
3) Category B, which refers to all other items.

Equipercentile Equating

An equipercentile equating method was used for equatings performed in this study. By definition, two scores from two different forms of one test may be considered equivalent to one another if their corresponding percentile ranks in any given group are equal (Kolen and Brennan, 2005). The relative cumulative frequency distribution for each form is computed and plotted. Examinees' scores are then matched for their equal percentile ranks. The single-group design was used in the equatings and re-equatings. In the single-group equating, the same group of examinees takes two forms of a test (Kolen and Brennan, 2005). In this study, the data from the same group was used in equating pretest scores to operational scores and in re-equating shortened pretest scores to the full pretest scores. Both the single group design and the equipercentile equating method are very straightforward. Smoothing was not used because the sample size in this research was sufficiently large.

Approximating Scaled Score Changes After Each Item Deletion

In this study, each focal-group frequency was used as the weighting factors in calculating the *STD FS-DIF* values for the 10 subgroups. A local linear approximation method (Dorans, 1984) was used to obtain the predicted scaled scores from the *STD FS-DIF* values for each of the 10 subgroups after the DIF items were deleted. The detailed derivations of the predicted scaled scores can be found in Appendix A. As can be seen in the estimation procedure, the predicted scaled score for each subgroup is a function of the *STD FS-DIF* index. It is this predicted scaled score that was used in testing the second hypothesis that the standardization method could predict the impact on the subgroup performance due to sizable DIF item deletion.

Results

In the SAT, DIF screening is a standard procedure for pretest items. Of the different pretests reviewed, only one was retained for this research because it had six C-level (sizable) DIF items. Three sizable DIF items were further analyzed in this research. It should be emphasized that none of these items was ever administered as an operational item on any SAT.

For DIF analyses, the matching variable was the operational score resulting from the 78-item verbal test. For the sake of simplicity, this test form will be referred to as Form X for the duration of this paper. Again, it

Table 2

Number of Examinees and Percentage of Total in the Data Sample

	African American	*Asian*	*Hispanic*	*White*	*All Others*	*Total*
Female	437 (4.6%)	299 (3.1%)	313 (3.3%)	3,799 (39.9%)	356 (3.7%)	5,204 (54.7%)
Male	345 (3.6%)	240 (2.5%)	229 (2.4%)	3,185 (33.5%)	314 (3.3%)	4,313 (45.3%)
Total	782 (8.2%)	539 (5.6%)	542 (5.7%)	6,984 (73.4%)	670 (7.0%)	9,517 (100%)

should be stated that the operational form of the SAT was adequately screened for DIF items given no C-level items. In total, there were 35 pretest items and 78 operational items on Form X.

The effects of deleting DIF items that are described in this study should be interpreted cautiously. First, the final forms of the SAT rarely contain DIF items because of the rigorous and proactive screening of pretests items. Second, the scaled scores used in this study were based on a particular pretest. It had only 35 items and was equated to the base, which was a 78-item operational test. The observed effects on this pretest resulted from the artificial circumstances associated with using a 35-item pretest to produce a test score. Dropping one item from a 78-item test affects scores much less than dropping one item from a 35-item test. It should be stated that we examined 60 pretests for DIF results before finding a pretest that had enough C items to adequately illustrate the dissection DIF approach.

Table 2 displays the number and percentages of examinees within gender and ethnic subgroups and the total group that received Form X. About three-quarters of the total sample were white examinees. Each of the four ethnic groups accounted for less than 9 percent of the total sample.

Effects of Deleting Items with C-Level DIF on Scaled Scores

Subgroup DIF analysis was performed on all items in the studied pretest form using the total operational score as the matching variable. The resulting Mantel-Haenszel DIF statistics provided information regarding which items exhibited sizable (C-level) DIF. Responses from these flagged items were then deleted from the computed raw scores. Three C-level DIF items, G1, E1, and G2, were selected for systematic item deletion. In total, there were three rounds of single-item deletion and one instance of removing all three items at once.

Dorans (1986) investigated the effects of item deletion on equating and scaling functions and reported scaled score distributions. He concluded that re-equating is psychometrically desirable after an item is deleted. In this research, equipercentile equating was used to equate the full pretest (35 items) to the operational test (78 items). Then, shortened tests (32 or 34 items, depending) were also equated to the operational test (78 items) using equipercentile equating. Re-equating using the equipercentile method was performed three times on the shortened 34-item test and once on the 32-item test (after removing items G1, E1, and G2 together). Resulting scaled scores were then compared between the full test and the shortened test forms. No smoothing was used since the sample was sufficiently large (n = 9,517). The distributions of the raw scores and scaled scores on a 20–80-point scale were obtained for each subgroup and total group. For this specific study, the scaled scores were expressed on a 20–80-point scale instead of a 200–800-point scale. The 20–80-point scale has one-scale point intervals.

Sample sizes and percentages by subgroup within its total group can be found in Table 3. It can be seen that females slightly outnumbered males in the total and each of the subgroups.

The one-way *STD FS-DIF* values and the two-way *STD FS-DIF* values for items G1, E1, and G2 can be found in subsequent tables. Initially using the one-way Standardization DIF procedure, items G1 and G2 were flagged for gender DIF effects, and item E1 was flagged for ethnic/racial DIF. The one-way *STD FS–DIF* values were derived from the traditional DIF analysis using males and whites as the reference groups. In contrast, the dissection *STD FS-DIF* values resulted from the two-way DIF methods using the total group as the reference group. Unrounded scaled score differences (SSDs) after removing each item are displayed as well.

As can be seen in Table 4, a one-way DIF procedure resulted in an *STD FS-DIF* index of –0.288 for item G1, using females as the focal group. The negative sign of this index indicates that the matched reference group (males) outperformed the focal group (females) and the matched white groups outperformed each of the three focal groups to which they are matched.

Using one-way DIF grouping, item G1 was flagged for gender DIF. In Table 5, the two-way *STD FS-DIF* indices distinctively show that, among male subgroups, white males had the largest positive DIF on item G1 (*STD FS-DIF* = 0.181). Among the female subgroups, African American females (*STD FS-DIF* = –0.202) and Asian females (*STD FS-DIF* = –0.192) had the most negative DIF. Other female subgroups (*STD FS-DIF* from –0.142 to –0.112) also had sizable negative DIF.

Comparing Table 5 to Table 4 reveals that the DIF on item G1 is mainly gender-based DIF. As shown in Table 5, all female subgroups had negative *STD FS-DIF* values, while all male subgroups had positive *STD FS-DIF* values. The *STD FS-DIF* differences across the female subgroups and across the male subgroups are trivial compared to the

STD FS-DIF differences between gender groups across the five ethnicities.

The *STD FS-DIF* effect on SSDs after dropping item G1 can be found in Table 6. On average, scaled scores (scale range 20–80) for all male subgroups were reduced, except for the Hispanic male group. The white male group lost 0.247 points. In contrast, each of the five female subgroups gained at least 0.152 points. For item G1, the groups that had the most negative DIF were the African American female and Asian female subgroups. On average, they gained the most: 0.297 and 0.266 points when item G1 was removed.

In Table 7, the one-way *STD FS-DIF* for male–female comparison was 0.012 (A-level DIF) for item E1. The one-way *STD FS-DIF* values were negative for all ethnic groups: –0.246 for African Americans, –0.165 for Asians, and –0.208 for Hispanics, showing that on item E1 the white group outperformed the individual ethnic groups to which it was matched. Thus, item E1 displays ethnic/racial DIF.

As seen in Table 8, the *STD FS-DIF* output resulting from the two-way scheme indicates that item E1 displays an ethnic/racial DIF effect between white subgroups and each individual ethnic/racial subgroup to which they were matched. The two-way *STD FS-DIF* values for white male and white female groups were 0.037 and 0.042, respectively, while all other ethnic/racial subgroups had negative DIF values. The African American females, Asian males, and Hispanic males had more negative DIF than the remaining subgroups.

The SSDs after dropping item E1 are indicated in Table 9. On average, the scaled score for the white female and white male groups decreased 0.084 and 0.056 points while the subgroups of African American, Asian, and Hispanic groups gained scores. Among the subgroups, the Asian males gained the most, 0.348 points on average, while the Asian female group gained the least, 0.080 points.

As seen in Table 10, the one-way DIF analysis results revealed that item G2 was another gender DIF item because male versus female *STD FS-DIF* was –0.193. Again, the results obtained by the two-way approach (Table 11) offer clarification at the subgroup level.

Values in Table 11 indicate that all male subgroups had positive *STD FS-DIF* values on item G2, while all female subgroups had negative *STD FS-DIF* values. African American females had the largest negative DIF of the female subgroups (*STD FS-DIF* = –0.146).

As seen in Table 12, after removing item G2, African American females, on average, gained the most points (0.166). Note also that they had the largest negative DIF in Table 11. The Asian male group, on the other hand, lost 0.224 points on average, followed by males in the all others category (–0.184), white males (–0.122), and Hispanic males (–0.121).

Table 13 summarizes the SSDs between the full pretest (35 items) and the shortened test (32 items) resulting from dropping items G1, E1, and G2. Among the ethnic/

Table 3

Numbers and Percentages of Males and Females Within Each Subgroup

	African American	*Asian*	*Hispanic*	*White*	*All Others*	*Total*
Female	437 (55.9%)	299 (55.5%)	313 (57.7%)	3,799 (54.4%)	356 (53.1%)	5,204 (54.7%)
Male	345 (44.1%)	240 (44.5%)	229 (42.3%)	3,185 (45.6%)	314 (46.9%)	4,313 (45.3%)
Total	782 (100%)	539 (100%)	542 (100%)	6,984 (100%)	670 (100%)	9,517 (100%)

Table 4

One-Way *STD FS-DIF* Values for Item G1

Reference/Focal group	STD FS-DIF
Male/Female	–0.288
White/African American	–0.140
White/Asian	–0.090
White/Hispanic	–0.087

Table 5

Two-Way *STD FS-DIF* Values for Item G1

	African American	*Asian*	*Hispanic*	*White*	*All Others*
Female	–0.202	–0.192	–0.142	–0.112	–0.132
Male	0.044	0.099	0.061	0.181	0.112

Table 6

Unrounded Scaled Score Differences (SSDs) After Removing Item G1

	African American	*Asian*	*Hispanic*	*White*	*All Others*
Female	0.297	0.266	0.152	0.161	0.166
Male	–0.034	–0.142	0.015	–0.247	–0.153

Table 7

One-Way *STD FS-DIF* Values for Item E1

Reference/Focal group	STD FS-DIF
Male/Female	0.012
White/African American	–0.246
White/Asian	–0.165
White/Hispanic	–0.208

Table 8

Two-Way *STD FS-DIF* Values for Item E1

	African American	*Asian*	*Hispanic*	*White*	*All Others*
Female	–0.166	–0.066	–0.124	0.042	–0.014
Male	–0.145	–0.176	–0.168	0.037	–0.035

Table 9

Unrounded Scaled Score Differences After Removing Item E1

	African American	*Asian*	*Hispanic*	*White*	*All Others*
Female	0.204	0.080	0.179	−0.084	−0.005
Male	0.315	0.348	0.271	−0.056	0.101

Table 10

One-Way *STD FS-DIF* Values for Item G2

Reference/Focal group	*STD FS -DIF*
Male/Female	−0.193
White/African American	−0.088
White/Asian	0.059
White/Hispanic	−0.008

Table 11

Two-Way *STD FS-DIF* Values for Item G2

	African American	*Asian*	*Hispanic*	*White*	*All Others*
Female	−0.146	−0.012	−0.077	−0.086	−0.076
Male	0.011	0.156	0.099	0.106	0.146

Table 12

Unrounded Scaled Score Differences After Removing Item G2

	African American	*Asian*	*Hispanic*	*White*	*All Others*
Female	0.166	−0.038	0.074	0.120	0.076
Male	0.024	−0.224	−0.121	−0.122	−0.184

Table 13

Unrounded Scaled Score Differences After Removing Items G1, E1, and G2

	African American	*Asian*	*Hispanic*	*White*	*All Others*
Female	0.782	0.391	0.490	0.191	0.255
Male	0.368	−0.091	0.098	−0.489	−0.302

racial subgroups, white males lost an average of 0.489 points, while scaled scores for African American females increased by an average of 0.782 points on a 20–80-point scale. There was a decrease of 0.302 for the all others male group mean and a slight decrease of 0.091 for the Asian male group mean. The remaining subgroups all had various degrees of increase of group means from the deletion of this set of items.

Summarized in Appendix A are the two-way *STD FS-DIF* effects for each of the flagged DIF items and unrounded observed scaled score differences (SSDs) for the 10 subgroups after DIF item deletion. They are also shown separately in Tables 5, 6, 8, 9, and 11 to 13. Items G1 and G2 were flagged for gender and item E1 was flagged for ethnic/racial DIF effects. It was generally the case that when a subgroup had a negative *STD FS-DIF* value on a certain item, their SSD was higher after removing that particular DIF item and re-equating. In other words, that subgroup gained more scaled score points when an item with negative DIF toward them was removed. For instance, the African American female group had the highest negative *STD FS-DIF* values on items G1 and G2. This group gained the most scaled score points when these two items were deleted one at a time. When all three items were removed simultaneously, the African American female group gained the most scaled score points among all subgroups. Although this was the typical case, it will not happen universally, especially if the magnitude of the DIF is close to zero. (For example, see Asian American female and item G2 in Appendix A.)

Prediction Based on the Standardization Approach

It was hypothesized that the effects of item deletion on scores could be predicted by the standardization DIF detection method. To be specific, the deletion of a negative DIF item should benefit the focal group, whereas the deletion of a positive DIF item should benefit the reference group. In order to test if the standardization method can indeed predict DIF effects of item deletion on scores, predicted SSDs on the shortened tests were obtained by applying the full test local linear approximation method (Dorans, 1984). Scaled score differences between the predicted and observed SSDs following item deletion and re-equating were obtained via the estimation process described in Appendix B. The estimation results of the predicted SSDs are summarized in Table 14.

As the results of both one-way and two-way *STD FS-DIF* indices indicated, items G1 and G2 displayed gender DIF, while item E1 displayed ethnic/racial DIF. The predicted SSDs in Table 14 show that all female groups gained score points, whereas male groups lost score points after removing items G1 or G2 from the full test. All ethnic/racial subgroup means increased, whereas the

Table 14

Predicted Scaled Score Differences for the Subgroups After DIF Item Deletion

Group	*Item G1*	*Item E1*	*Item G2*
African American Female	0.267	0.219	0.193
Asian Female	0.278	0.096	0.017
Hispanic Female	0.204	0.178	0.111
White Female	0.168	−0.063	0.129
All Other Female	0.198	0.021	0.114
African American Male	−0.072	0.236	−0.018
Asian Male	−0.143	0.255	−0.226
Hispanic Male	−0.079	0.217	−0.128
White Male	−0.262	−0.054	−0.154
All Other Male	−0.162	0.051	−0.211

white female and white male group means decreased when item E1 was removed.

Correlation analyses were conducted between predicted SSDs and observed SSDs for each subgroup after each item deletion (see Table 15).

It can be seen that the correlation coefficients between predicted SSDs and observed SSDs ranged from 0.97 to 0.98, close to 1, a perfect positive correlation. These high positive correlations highlight the strong positive relationship between predicted SSDs and the observed SSDs. When predicted SSD increases (i.e., the focal group benefits from the item deletion), the observed SSD value for that item also increases. When predicted SSD decreases (i.e., focal group is disadvantaged by the item deletion), the observed SSD value for that item also decreases. All other deletions of an item with negative DIF result in a negative change in observed SSDs.

Scatterplots between mean predicted SSDs and the observed SSDs after removing items G1, E1, and G2 are shown in Figures 1–3. These scatterplots clearly depict the strong positive relationship between the predicted SSDs and the observed SSDs after each sizable DIF item was removed. Numbers 1 through 10 in the scatterplots represent group membership, where 1 = African American Females, 2 = Asian Females, 3 = Hispanic Females, 4 = White Females, 5 = All Other Females, 6 = African American Males, 7 = Asian Males, 8 = Hispanic Males, 9 = White Males, and 10 = All Other Males.

Table 15

Correlation Between Predicted SSDs and Observed SSDs

Item Deleted	*Correlation*
G1	0.98
E1	0.97
G2	0.98

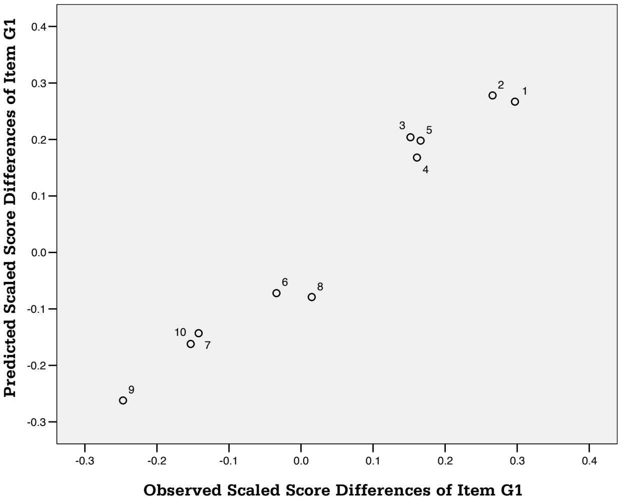

Figure 1. Scatterplot of predicted mean SSDs and observed mean SSDs after removing item G1.

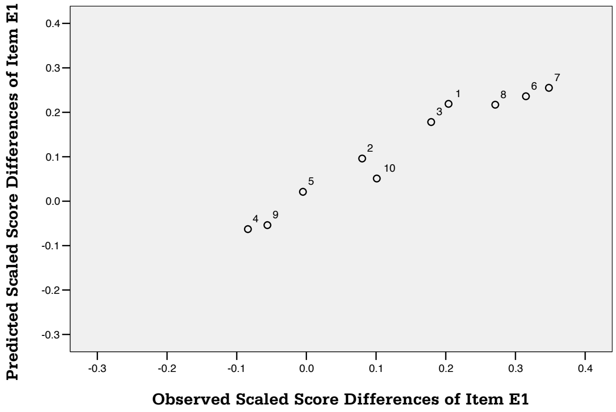

Figure 2. Scatterplot of predicted mean SSDs and observed mean SSDs after removing item E1.

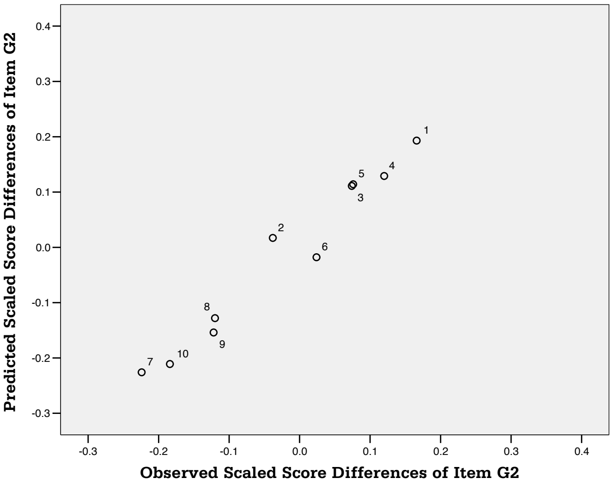

Figure 3. Scatterplot of predicted mean SSDs and observed mean SSDs after removing item G2.

Discussion

This research has shown that the act of deleting large DIF items from an assessment instrument can differentially affect subgroup-level performance. In this research, the reference group was defined to be the total group, while each of the subgroups independently acted as a focal group. We call this the DIF dissection method. Since different DIF effects exist in each subgroup, it is believed that using a combination of all groups as the reference group and each combination of gender and ethnicity as a focal group produces more accurate, though potentially less stable, findings than using a simple majority group approach. Groups defined by crossing gender and ethnicity will have smaller samples than those that aggregate across gender or ethnicity. Small samples mean less stability. On the other hand, finer definitions of groups will lead to more accurate estimates of DIF effects for these finer defined groups. For example, neither the African American DIF estimates nor the female DIF estimate are likely to be as accurate for African American females as the African American female DIF estimates would be.

As hypothesized, the deletion of a sizable DIF item most disadvantageous to a particular group has been shown to affect the scores of that group the most. Scaled score differences after item deletion and re-equating varied among subgroups depending on the DIF effects. Those groups found to be disadvantaged via the two-way DIF approaches when all three items were deleted gained points, whereas those thought to be advantaged lost points. In particular, African American females gained the most when all three items were deleted, which was consistent with the fact that they were disadvantaged on all of the items. However, the overall gained and lost points amounted to less than one scale-point on a 20–80-point scale.

It was also hypothesized that the effects of item deletion on scores can be predicted based on the standardization method. This hypothesis was tested by obtaining predicted scaled scores on the shortened tests via applying the full test local linear approximation. Correlation and scatterplots confirmed that the standardization DIF method could reliably predict score changes.

The DIF dissection method and two-way classification method may benefit large-scale standardized testing programs. The DIF dissection method places everyone in the reference group simultaneously. The purpose of using the DIF dissection within the context of a two-way classification procedure is to examine gender by ethnicity interactions that traditional DIF grouping methods, i.e., one-way methods, do not allow. This more informative approach to DIF analysis not only confirms findings from the traditional (one-way) DIF approach, but also enhances our understanding of the behavior of DIF items. It was shown that the act of deleting a large DIF item can (and does) have differential impact at the subgroup level. DIF detection procedures done via a two-way approach can offer valuable help to the decision-making process, especially when determining impact due to item deletion prior to score reporting. Additional information can be obtained by looking at the scaled score changes at the subgroup level and proactively surveying to what extent the most disadvantaged groups may be affected.

One way to understand the difference between the one-way DIF analysis and the two-way DIF method is through the analogy of analysis of variance (ANOVA). In terms of research design, conducting a one-way DIF analysis is similar to conducting a one-way ANOVA, where each ethnic/racial group and gender group functions as a main effect. In contrast, a two-way DIF analysis is similar to a two-way ANOVA, where information regarding interactions is available.

A limitation of this study was the limited sample sizes for ethnic/racial subgroups. In cases where small samples are used for analyses, the standardization method might produce unstable DIF estimates and prevent generalization of the results. A possible follow-up study to this research could be to apply kernel smoothing, a process currently used in the ETS comprehensive statistical analysis system GENASYS. This process is usually reserved for total group analyses only. One possibility is to investigate using kernel smoothing on small samples so as to facilitate subgroup DIF analyses. A common standardization group approach can also be used to obtain *STD FS-DIF* values and then compare their impact on subgroup-level DIF effects and scaled score changes.

References

Carlton, S. T., & Harris, A. M. (1992). *Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons*. Princeton, NJ: Educational Testing Service.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement*, *24*, 157–66.

Dorans, N. J. (1984). *Approximate IRT formula score and scaled score standard error of measurements at different ability levels* (SR-84-118). Princeton, NJ: Educational Testing Service.

Dorans, N. J. (1986). The impact of item deletion on equating conversions and reported score distributions. J*ournal of Educational Measurement*, *23*(3), 245–64.

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

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281–306.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–68.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In Bennett, R. E. & Ward, W. C. (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Holland, P. W., & Thayer, D. T. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129–45). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hu, P. G., & Dorans, N. J. (1989). *The effect of deleting differentially functioning items on equating functions and reported score distributions*. Princeton, NJ: Educational Testing Service.

Kolen, M. J., & Brennan, R. L. (2005). *Testing equating, scaling and linking*. New York: Springer-Verlag.

O'Neil, 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. 255–76). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and African American examinees. Applied Measurement in Education, *10*(4), 299–319.

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

Zhang, Y. (2001). Differential item functioning in a large scale mathematics assessment: The interaction of gender and ethnicity. Unpublished dissertation, Ohio University.

Zieky, M. J. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–47). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Appendix A: Summary of Two-Way *STD FS-DIF* Effects and Unrounded Observed Scaled Score Differences for the Subgroups After DIF Item Deletion

	Two-Way STD FS-DIF *Effects*			*Unrounded Observed SSDs After DIF Item Deletion*			
Group	***Item G1***	***Item E1***	***Item G2***	***Item G1***	***Item E1***	***Item G2***	***All 3 Items***
African American Female	−0.202	−0.166	−0.146	0.297	0.204	0.166	0.782
Asian Female	−0.192	−0.066	−0.012	0.266	0.080	−0.038	0.391
Hispanic Female	−0.142	−0.124	−0.077	0.152	0.179	0.074	0.490
White Female	−0.112	0.042	−0.086	0.161	−0.084	0.120	0.191
All Other Female	−0.132	−0.014	−0.076	0.166	−0.005	0.076	0.255
African American Male	0.044	−0.145	0.011	−0.034	0.315	0.024	0.368
Asian Male	0.099	−0.176	0.156	−0.142	0.348	−0.224	−0.091
Hispanic Male	0.061	−0.168	0.099	0.015	0.271	−0.121	0.098
White Male	0.181	0.037	0.106	−0.247	−0.056	−0.122	−0.489
All Other Male	0.122	−0.035	0.146	−0.153	0.101	−0.184	−0.302

Appendix B: The Local Approximation Algorithm

Dorans (1984) developed an algorithm for obtaining predicted scaled scores via local slopes and intercepts. The details for the portion of that estimation procedure that pertain to this study are described below.

The conversion table that converts formula scores of the full pretest to scaled scores (in intervals of 1) is searched for four values. The table below is a portion of the conversion table.

Pretest Raw Score	*Unrounded Scaled Score*
–9	
—	—
0	29.56
—	—
5	38.19
6	39.85
7	**41.30**
8	**42.62**
9	44.25
10	45.74
—	—
—	—
35	

Take the African American Female group on item G1 for an example. The unrounded mean scaled score of the group is 42.37 (SSj). Looking in the unrounded scaled score column, we can see that

U(SSj)—the upper scaled score closest to 42.37 was 42.62
U(FSj)—the corresponding raw formula score to U(SSj) was 8
L(SSj)—the lower scaled score closest to 42.37 was 41.30
L(FSj)—the corresponding raw formula score to L(SSj) was 7

Then, a local approximation to the slope of the raw-to-scale conversion near SSj can be obtained via

$$\mathrm{Aj} = \frac{\mathrm{U(SSj)} - \mathrm{L(SSj)}}{\mathrm{U(FSj)} - \mathrm{L(FSj)}} = \frac{42.62 - 41.30}{8 - 7} = 1.32$$

Apply this local slope to the negative of the two-way DIF effect size to get a predicted scaled score. For example, suppose the *STD FS-DIF* effect size is –.202, and the slope at the focal group mean is 1.32, then the predicted effect is 1.32*(–)(–.202) = 0.267. This value was the predicted scaled score difference for African American female group after item G1 was removed (see Table 14). These predicted changes in mean scaled scores were compared to the changes in mean scaled score values resulted from the re-equating after deletion of a sizable DIF item.

CollegeBoard
connect to college success™

www.collegeboard.com