ED 476 147	TM 034 896
AUTHOR	Zhang, Yanling; Matthews-Lopez, Joy; Dorans, Neil J.
TITLE	Using DIF Dissection To Assess Effects of Item Deletion Due to DIF on the Performance of SAT I: Reasoning Test Sub-populations.
INSTITUTION	Educational Testing Service, Princeton, NJ.
PUB DATE	2003-04-00
NOTE	35p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Classification; College Entrance Examinations; High School Students; High Schools; *Item Bias; *Scores; *Test Items
IDENTIFIERS	Item Deletion

ABSTRACT

Statistical procedures for detecting differential item functioning (DIF) are often used to screen items for construct irrelevant variance. Standard DIF detection procedures focus on only one categorical variables at an aggregated group or one-way level, like gender or ethnicity/race. Building on previous work by P. Hu and N. Dorans (1998), N. Dorans and P. Holland (1993), and Y. Zhang (2001), this study applied a DIF dissection classification scheme to SAT:I Verbal data. Subsequently, the effect of deleting sizable DIF items on reported scores after equipercentile re-equating were explored using data from a spring administration of the SAT for 9,517 test takers in 10 subgroups. By using a "dissection" approach to reference and focal group formations, this two-way classification scheme may yield new and detailed insight into item functioning at the subgroup level. Two hypotheses were studied: (1) whether or not the deletion of sizeable DIF items disadvantageous to a particular subgroup will affect 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 were predicted by the results of this research. Scaled score differences following item deletion and re-equating varied among subgroups, depending on the DIF effects. Subgroups disadvantaged by the subsequently deleted sizable DIF items gained scaled score points whereas advantaged groups lost. Regression analyses confirmed the second hypothesis. It was also shown that by deleting an item with sizable negative DIF, the focal group might be greatly benefited. Among three item deletion scenarios, DIF effects yielded from the two-way classification scheme showed very little interaction in the majority of cases. (Contains 3 figures, 14 tables, and 15 references.) (Author/SLD)

ETS Educational Testing Service

Using DIF Dissection to Assess Effects of Item Deletion due to DIF on the Performance of SAT® I: Reasoning Test Sub-populations

Yanling Zhang

Joy Matthews-López

Neil J. Dorans

Educational Testing Service

Princeton, New Jersey

Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) held between April 21 to 25, 2003, in Chicago, IL.

Educational Testing Service
Princeton, New Jersey 08541

Abstract

Statistical procedures for detecting differential item functioning (DIF) are often used to screen items for construct irrelevant variance. Standard DIF detection procedures focus on only one categorical variable at an aggregated group or one-way level, like gender or ethnicity/race. Building on previous work by Hu and Dorans (1989), Dorans and Holland (1993), and Zhang (2001), this research applies a DIF dissection classification scheme to SAT I: Verbal data. Subsequently, the effect of deleting sizable DIF items on reported scores after equipercentile re-equating were explored. By using a "Dissection" approach to reference and focal group formations, this two-way classification scheme may yield new and detailed insight into item functioning at the subgroup level. Two hypotheses were studied: (1) whether or not the deletion of sizeable DIF items disadvantageous to a particular subgroup will affect 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 were supported by the results of this research. Scaled score differences following item deletion and re-equating varied among subgroups, depending on the DIF effects. Subgroups disadvantaged by the subsequently deleted sizable DIF items gained scaled score points whereas those advantaged, lost. Regression analyses confirmed the second hypothesis. It was also shown that by deleting an item with sizable negative DIF, the focal group might be greatly benefited. Among three item deletion scenarios, DIF effects yielded from the two-way classification scheme showed very little interaction in the majority of cases.

Background

Standardized achievement tests often have high stakes attached to their use. Statistical procedures for detecting differential item functioning (DIF) are frequently used to screen items for construct irrelevant variance. 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., Calton & Harris, 1992; Doolittle & Cleary, 1987; O'Neil & McPeek, 1993; Scheuneman & Grima, 1997; and Schmitt & Dorans, 1990).

Hu and Dorans (1989) used data from the SAT I: Verbal test 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 or not the deletion of minimal and/or sizable DIF items resulted in different scaled scores after IRT true score re-equating and Tucker re-equating. The results of that study indicated that though deleting certain items affected scaled scores in general, the act of deleting the item itself had a larger effect on scaled scores than did the extent of DIF of the deleted items.

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" or "Dissection" 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 two-way DIF classification scheme to SAT® I: Reasoning Test (SAT): Verbal data. Subsequently, the effect of deleting sizable DIF items on reported scores after equipercentile re-equating was examined. As mentioned earlier, this two-way classification scheme utilizes non-traditional 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. It is believed that using this approach to reference and focal group formation may provide detailed information about item performance at the subgroup-level.

There were three goals to this research: (1) to examine items for DIF using the above-described DIF dissection classification scheme within the standardization DIF detection procedure, (2) to assess the effect, if any, of deleting sizable DIF items from all groups on the reported score after re-equating the shortened tests, and (3) to make recommendations regarding future routine DIF detection procedures.

The hypotheses to be tested are the following:

1. The deletion of DIF items disadvantageous to a particular subgroup will affect that subgroup the most;
2. The effects of item deletion on scores will be predicted by the standardization method.

All items of a particular SAT I: Verbal pretest that were flagged for sizable levels of DIF during standard operational analyses were removed from the response vectors of the affected group as well as from all other groups. Sizable DIF is defined according to the ETS delta criteria and will be elaborated upon more in the later part of this report. Reported score distribution and score changes of each ethnic and gender group were then examined after the systematic deletion of each item. The standardization method (Dorans & Kulick, 1986) was chosen for this work because it is easily adapted to formula scored test as well as to the scenario of multi-group analyses. It also lends itself well to the prediction of effects of item deletion on subgroup performance.

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, non-operational section that was used for equating purposes as well as for pretesting new items. This research is limited to the use of 35 five-option multiple-choice verbal pretest sections. 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), we placed all examinees who indicated their gender but not their ethnicity/race in a group labeled as "All Others." In addition, Native Americans were also placed in "All Others" since this particular sample size was too small to withstand subgroup-level analyses.

A total of ten 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.)

Table 1

Composition of Reference Group and Focal Groups

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

Formula Scoring Procedures

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

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

where k=number of options for each multiple-choice item. As can be seen, omitted and not reached items (NR) are treated differently than are incorrect responses. Whereas points are neither awarded nor deducted for omitted/not reached items, incorrect responses to the multiple-choice result in the loss of a fraction of a point. In this case, 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 & Schmitt, 1993) was used in this study. As stated by Dorans and Holland (1993), standardization method is readily adopted to a formula-scored item, such as those used on SAT-Verbal.

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

$$FS_{STD} = \frac{\sum_{m=1}^{M}\left[K_m\left(FS_{fm} - FS_{rm}\right)\right]}{\sum_{m=1}^{M} K_m},$$

where $\frac{Km}{\sum_{m=1}^{M} Km}$ is the weighting factor at score level *m*. Score level *m* is supplied by the standardization group to weight differences in item performance between the focal group, *FSfm,* and the reference group, *FSrm.*

Since the SAT is a formula-scored test, formula-scored DIF given by the standardization method indices, *STD FS-DIF*, is used for DIF 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* scores item as 1 if correct and 0 if incorrect, 0 if omitted, or 0 if NR, it incorporates a formula scoring algorithm and assigns zero weight to omitted and NR items, and $[\frac{-1}{k-1}]$ to incorrect responses, where k is the number of choice options. The *STD FS-DIF* index ranges between –1.25 to +1.25, inclusive in this case where k=5.

ETS Classification Criteria

Educational Testing Service (ETS) relies on a DIF statistic that expresses differences on a delta scale as a measure of magnitude of effect. In order to compute this statistic, the Mantel-Haenszel common-odds-ratio, α_{MH}, must first be computed. After α_{MH} is derived, it is placed on a delta scale via the following logarithmic transformation (Holland & Thayer, 1988): delta$_{MH}$ = -

2.35 ln(α_{MH}). Delta$_{MH}$ can be interpreted as the average amount a member of the reference group found the studied item to be more difficult than did a comparable group member of the focal group, or vice versa. A value of zero suggests no DIF is present. Similar to the *STD FS-DIF* index, a negative DIF value suggests that that the focal group is disadvantaged and the reference group is advantaged while a positive DIF value indicates that focal group is advantaged and the reference group is disadvantaged.

Dorans and Holland (1993) described the ETS DIF classification scheme for use in test development. According to the scheme, all the items can be categorized into one of following three non-overlapping groups:

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

Equipercentile Equating

An equipercentile equating method was used for the equating 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, 1988). The relative cumulative frequency distribution for each form is computed and plotted. Examinees scores are then matched for their equal percentile ranks. Both the single group design and the equipercentile equating method are very straightforward. Smoothing was not needed since the sample size in this research was sufficiently large.

Results

DIF summary statistics were reviewed 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 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. 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 used to compute delta$_{MH}$ 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 should be stated that the operational form of the SAT was DIF-free since no C-level items are ever used on operational test forms. In total, there were 35 pretest items and 78 operational items on Form-X.

Table 2 displays the number of examinees and percentages of subgroups out of the grand total group which received Form-X.

Table 2

Number of Examinees and Percentage of Total in the Data Sample

	African American	Asian	Hispanic	White	All Others	Total
Female	437 (4.59%)	299 (3.14%)	313 (3.29%)	3,799 (39.92%)	356 (3.74%)	5,204 (54.68%)
Male	345 (3.63%)	240 (2.52%)	229 (2.41%)	3,185 (33.47%)	314 (3.30%)	4,313 (45.32%)
Total	782 (8.22%)	539 (5.66%)	542 (5.70%)	6,984 (73.38%)	670 (7.04%)	9,517 (100%)

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

Subgroup DIF analysis was performed on all items in the studied pretest using the operational score as the matching variable. The resulting 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, Item #1, Item #11, and Item #16 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/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. No smoothing was needed since the sample was sufficiently large (n=9,517). A standard formula scoring procedure was used as discussed earlier in this paper. The distributions of the raw scores and scaled scores on a 20-80 scale were obtained for each subgroup and total group. For this specific study, the scaled scores were expressed on a 20 to 80 point scale instead of a 200-800 scale so that the observed SSDs could be expressed in perspective.

Re-equating using the equipercentile method was then performed: three times on the shortened 34-item test and once on the 32-item test (after removing items #1, #11, and #16, together). Resulting scaled scores were then compared between the full test and the shortened test forms.

Table 3

Numbers and Percentages of Males and Females within each Subgroup

	African American	Asian	Hispanic	White	All Others	Row Total
Female	437 (55.88%)	299 (55.47%)	313 (57.75%)	3,799 (54.40%)	356 (53.13%)	5,204 (54.68%)
Male	345 (44.12%)	240 (44.53%)	229 (42.25%)	3,185 (45.60%)	314 (46.87%)	4,313 (45.32%)
Column Total	782 (100%)	539 (100%)	542 (100%)	6,984 (100%)	670 (100%)	9,517 (100%)

Sample sizes and percentages by subgroup within its total group can be found in Table 3. The one-way STD FS-DIF values and the two-way STD FS-DIF for items #1, #11, and #16 can be found in subsequent tables. The one-way STD FS –DIF values were derived from the traditional DIF analysis using the 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.

Table 4

One-way STD FS –DIF Values for Item #1

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

As seen in Table 4, a one-way DIF procedure resulted in a *STD FS-DIF* index of -0.288 for Item #1 (using females as focal group). The negative sign of this index indicates that the reference group (males) outperformed the focal group (females), suggesting that this item disadvantaged the female group.

In Table 5 below, the two-way *STD FS-DIF* indices distinctively show that, among male subgroups, White males benefited most from Item 1 (*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) were most adversely affected, though other female subgroups (*STD FS-DIF* from –0.142 to –0.112) were negatively affected as well.

Table 5

The Two-way STD FS –DIF Values for Item #1

	African American	Asian	Hispanic	White	All Others	Total
Female	-0.202	-0.192	-0.142	-0.112	-0.132	-0.127
Male	0.044	0.099	0.061	0.181	0.112	0.154
F – $M_{Difference}$	-0.246	-0.291	-0.203	-0.293	-0.244	-0.281
Total	-0.093	-0.062	-0.056	0.021	-0.017	---

Row three in Table 5 displays the difference between the female and male two-way *STD FS–DIF* values for each ethnic group. The values are not significantly different from each other, ranging from –0.203 to –0.291, thus, showing little gender by ethnicity interaction.

Table 6

Unrounded Scaled Score Differences after Removing Item #1

	African American	Asian	Hispanic	White	All Others	Total
Female	0.327	0.328	0.195	0.206	0.211	0.223
Male	-0.015	-0.163	0.066	-0.198	-0.096	-0.160
Total	0.176	0.110	0.140	0.022	0.067	0.049

The *STD FS-DIF* effect on the scaled score differences (SSDs) after dropping Item #1 can be seen 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 0.198 points. In contrast, on average, each of the five female groups gained at least 0.195 points. For Item #1, the groups that were most seriously affected by the DIF were African American female and Asian female subgroups. On average, they gained most: 0.327 and 0.328 points when Item #1 was removed.

Table 7

One-way STD FS –DIF Values for Item #11

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

In Table 7, the one-way *STD FS-DIF* for male/female comparison was 0.012 (A-level DIF). In Table 8, the two-way *STD FS-DIF* output resulting from the two-way scheme indicates that Item #11 displays a DIF effect between White and each individual ethnic group. 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. Item #11 gave a slight advantage to the White group over individual ethnic groups; the two-way *STD FS-DIF* values (Table 8) for White male and female groups were 0.037 and 0.042, respectively. It should be noted that African American females, Asian males and Hispanic males were more seriously affected than the remaining subgroups.

Table 8

The Two-way STD FS –DIF Values for Item #11

	African American	Asian	Hispanic	White	All Others	Total
Female	-0.166	-0.066	-0.124	0.042	-0.014	0.005
Male	-0.145	-0.176	-0.168	0.037	-0.035	-0.006
F – $M_{\text{Difference}}$	-0.021	0.11	0.044	0.005	0.021	0.011
Total	-0.157	-0.115	-0.142	0.040	-0.024	---

In Table 8, the female and male difference for the Asian group is 0.11 while other groups were greatly lower than 0.044. The DIF effect for Asian males was more than twice as much compared to the Asian females (-0.176 vs. –0.066), thus, showing gender by ethnicity interaction.

Table 9

Unrounded Scaled Score Differences after Removing Item #11

	African American	Asian	Hispanic	White	All Others	Total
Female	0.121	0.017	0.128	-0.110	-0.037	-0.065
Male	0.203	0.288	0.205	-0.083	0.064	-0.013
Total	0.151	0.137	0.161	-0.098	0.011	-0.042

The SSDs after dropping Item #11 are indicated in Table 9. On average, the scaled score for the White group, as a whole, decreased 0.098 points while African American, Asian, and Hispanic groups gained 0.151, 0.137, and 0.161 points, respectively. By inspecting subgroups, it can be seen that the Asian males gained most, 0.288 points on average, followed by 0.205 points for Hispanic males and 0.203 points for African American males. African American males were also the most disadvantaged subgroup, as indicated by the two-way *STD FS-DIF* values seen in Table 8.

Table 10

One-way STD FS –DIF Values for Item #16

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

As indicated in Table 10, the one-way DIF analysis results revealed that Item #16 was another gender DIF item (*STD FS-DIF* = –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 African American females were the most disadvantaged of the female subgroups (*STD FS-DIF* = -0.146), as seen in Table 11. All male subgroups yielded positive *STD FS-DIF* values.

Table 11

The Two-way STD FS –DIF Values for Item #16

	African American	Asian	Hispanic	White	All Others	Total
Female	-0.146	-0.012	-0.077	-0.086	-0.076	-0.086
Male	0.011	0.156	0.099	0.106	0.146	0.103
F – $M_{Difference}$	-0.157	-0.168	-0.176	-0.192	-0.222	-0.189
Total	-0.077	0.063	-0.003	0.001	0.028	---

Table 11 indicates that after deleting item #16, the female and male DIF effects differ similarly (from -0.157 to -0.222), showing little gender by ethnicity interaction.

Table 12

Unrounded Scaled Score Differences after Removing Item #16

	African American	Asian	Hispanic	White	All Others	Total
Female	0.144	-0.060	0.032	0.131	0.082	0.112
Male	0.012	-0.221	-0.135	-0.113	-0.166	-0.114
Total	0.086	-0.139	-0.039	0.020	-0.034	0.009

As seen in Table 12, after removing Item #16, African American females, on average, gained the most points (0.144); note also that they were the most disadvantaged group shown in Table 11. The Asian male group, on the other hand, lost 0.221 points on average, followed by All Others males (0.166), Hispanic males (0.135), and White males (0.113).

Table 13

Unrounded Scaled Score Differences after Removing Item #1, #11, and #16

	African American	Asian	Hispanic	White	All Others	Total
Female	0.789	0.355	0.482	0.173	0.244	0.259
Male	0.354	-0.150	0.101	-0.517	-0.296	-0.378
Total	0.597	0.130	0.321	-0.142	-0.009	-0.030

Table 13 summarizes the SSDs between the full pretest (35 items) and the shortened test (32 items) resulting from dropping Item #1, Item #11, and Item #16. Between male and female groups, males lost an average of 0.378 scaled points and females gained, on average, 0.259 points after dropping all three items. Among the intact ethnic groups, the average score increase for African Americans was 0.597 points, followed by an increase of 0.321 points for the Hispanic group. Within subgroups, White males lost an average of 0.517 points while scaled scores for African American females increased by an average of 0.789 points. Hispanic females also gained an average of 0.482 points from the deletion of this set of items.

Obtaining the One-way DIF Using Subgroup Two-way DIF

Two-way DIF indices for males and females within each subgroup can be used to derive the one-way DIF indices for each gender and ethnic group. When two subpopulations are of equal size, the total population (TP) average, indicated by AveTP, is simply the weighted sum of the two sub-population (SP) averages, referred to as AveSP1 and AveSP2, respectively:

$$\text{AveTP} = 0.5^{*}\ (\text{AveSP1} + \text{AveSP2}) \qquad (1)$$

The difference between AveSP1 and AveTP is defined to be

$$\text{AveSP1} - \text{AveTP} = \text{AveSP1} - .5^{*}(\text{AveSP1} + \text{AveSP2})$$

$$\Rightarrow \quad \text{AveSP1} - \text{AveTP} = .5^{*}(\text{AveSP1} - \text{AveSP2}).$$

By the same reasoning,

$$\text{AveSP2} - \text{AveTP} = \text{AveSP2} - .5^{*}(\text{AveSP1} + \text{AveSP2})$$

$$\Rightarrow \quad \text{AveSP2} - \text{AveTP} = .5^{*}(\text{AveSP2} - \text{AveSP1}).$$

However, generally, sub-populations have unequal sample sizes. When two subpopulations are of unequal sizes, the total population AveTP is simply the weighted sum of the two sub-population averages, AveSP1 and AveSP2, respectively, in which the weights sum to 1.

Hence,
$$\text{AveTP} = (\text{w1}^*\text{AveSP1} + \text{w2}^*\text{AveSP2}), \qquad (2)$$
where weights w1 and w2 are the proportions of sample sizes for each subgroup, and w1 + w2 =1.

In the context of regular one-way DIF analysis, the object of investigation is the difference between two total group means, for example, male and female groups, which is equivalent to (AveSP1 – AveSP2). However in the context of two-way DIF, the differences between each subgroup and the total group mean is examined, which is equivalent to (AveSP1-AveTP).

Therefore,
$$\text{AveSP1} - \text{AveTP} = \text{AveSP1} - (\text{w1}^*\text{AveSP1} + \text{w2}^*\text{AveSP2})$$
$$\Rightarrow \quad \text{AveSP1} - \text{AveTP} = \text{w2}^*(\text{AveSP1} - \text{AveSP2}) \qquad (3)$$

and

$$\text{AveSP2} - \text{AveTP} = \text{AveSP2} - (\text{w1}^*\text{AveSP1} + \text{w2}^*\text{AveSP2})$$
$$\Rightarrow \quad \text{AveSP2} - \text{AveTP} = \text{w1}^*(\text{AveSP2} - \text{AveSP1}) \qquad (4)$$

In the context of this paper, we make the assumption that AveTP or Ave DIF in the total population = 0. Thus, equation (3) becomes,

$$\text{AveSP1} = \text{w2}^*(\text{AveSP1} - \text{AveSP2}), \qquad (5)$$

and equation (4) becomes,

$$\text{AveSP2} = \text{w1}^*(\text{AveSP2} - \text{AveSP1}) = -\text{w1}^*(\text{AveSP1} - \text{AveSP2}) \qquad (6)$$

In the context of one-way DIF procedures, the one-way male vs. female DIF is (AveSP1-AveSP2). By simple algebraic manipulations, the difference of equations (5) and (6) can be expressed as

$$\text{AveSP1-AveSP2} = \text{AveSP1/w2} = \text{-AveSP2/w1}. \qquad (7)$$

By applying equation (2), the one-way DIF indices for ethnic groups can be obtained by summing the weighted ethnic subgroup DIF indices, where the weights being applied are the proportion of their sample sizes over the total group as shown in Table 3. Consider the following example for Item #1.

Example 1:

Let SP1 be African American females and SP2 be African American males. Equation (2) can be expressed as $\text{AveTP}_{\text{AF}} = (\text{w1*AveSP1}_{\text{female}} + \text{w2*AveSP2}_{\text{male}})$. The weight for subgroup 1 (African Americans females) is .5588 and the weight for subgroup 2 (African Americans males) is .4412; these values can be found in Table 3. In Table 5, the two-way *STD FS-DIF* values for $\text{AveSP1}_{\text{female}}$ and $\text{AveSP2}_{\text{male}}$ are –0.202 and 0.044, respectively. By substituting these values into equation (2), we derive the following value:

$\text{AveTP}_{\text{AF}} = (0.5588*(-0.202) + 0.4412*0.044) = (-0.11288) + 0.01941 = -0.09347$,

which is the one-way DIF value for the African American group (see in Table 5). It can be shown that one-way DIF values for all others ethnic and gender groups can be derived accordingly.

Equation (7) can also be used to obtain one-way gender DIF values on Item #1. Consider the following example.

Example 2:

Let SP1 be Females and SP2 be Males. $\text{AveSP1}_{\text{female}}$ and $\text{AveSP2}_{\text{male}}$ can be found from the two-way DIF values located in Table 6 and proportions of male and female groups located in Table 4. Equation (5) for this Male/Female DIF is:

$$\text{AveSP1}_{\text{female}} - \text{AveSP2}_{\text{male}} = \text{AveSP1}_{\text{female}}/w_{\text{male}} = (-0.127)/(.4532) = -0.280,$$

which is close to the one-way DIF value (-0.288) for Male/Female DIF (see Table 5.) Also note that

$$\text{AveSP1}_{\text{female}} - \text{AveSP2}_{\text{male}} = -\text{AveSP2}_{\text{male}}/w_{\text{female}} = -(0.154)/(.5468) = -0.282$$

is remarkably close to the one-way DIF value (-0.288) for Male/Female DIF shown in Table 4.

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, correlation analyses were conducted between SSDs for each subgroup after each item deletion scheme and their formula score DIF effects.

Table 14

Correlation Between Scaled Score Differences and STD FS-DIF Indices

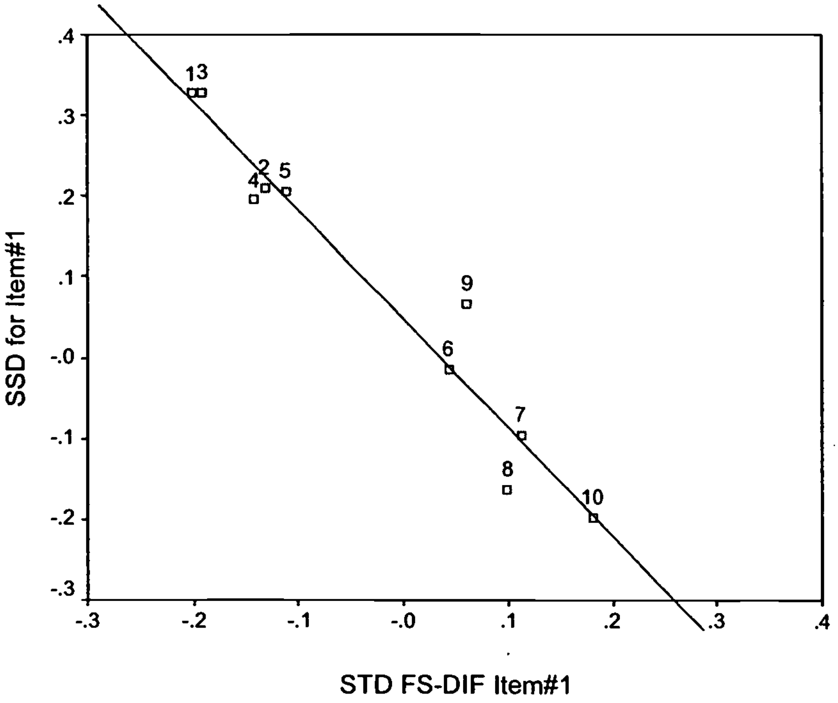
Item Deleted	Correlation
#1	-0.972
#11	-0.944
#16	-0.963

Table 14 shows that the correlation indices were very strongly negative. These high negative correlations highlight the strong negative relationship between SSDs and the two-way *STD FS-DIF* indices. When SSD increases (i.e., the focal group benefits from the item deletion), the two-way *FS STD-DIF* value for that item is negative. When SSD decreases (i.e., focal group is disadvantaged by the item deletion), the two-way *FS STD-DIF* value for that item is positive. All other deletions of an item with negative DIF result in a positive change in scaled scores.

Linear regression analyses were performed between SSDs following the removal of each item and two-way *STD FS-DIF* values (the predictor variable). Scatter plots between mean SSDs and the two-way DIF indices after removing Items #1, #11, and #16, are shown in Figures 1-3. Numbers 1 through 10 in the scatter plots represent group membership, where 1=African American Females, 2=All Others Females, 3=Asian Females, 4=Hispanic Females, 5=White Females, 6=African American Males, 7=All Others Males, 8=Asian Males, 9=Hispanic Males, and 10=White Males. Corresponding regression equations appear below each figure. It should be stated that these regression equations were included as they were descriptive for this small sample. They have little generalizability.

Figure 1

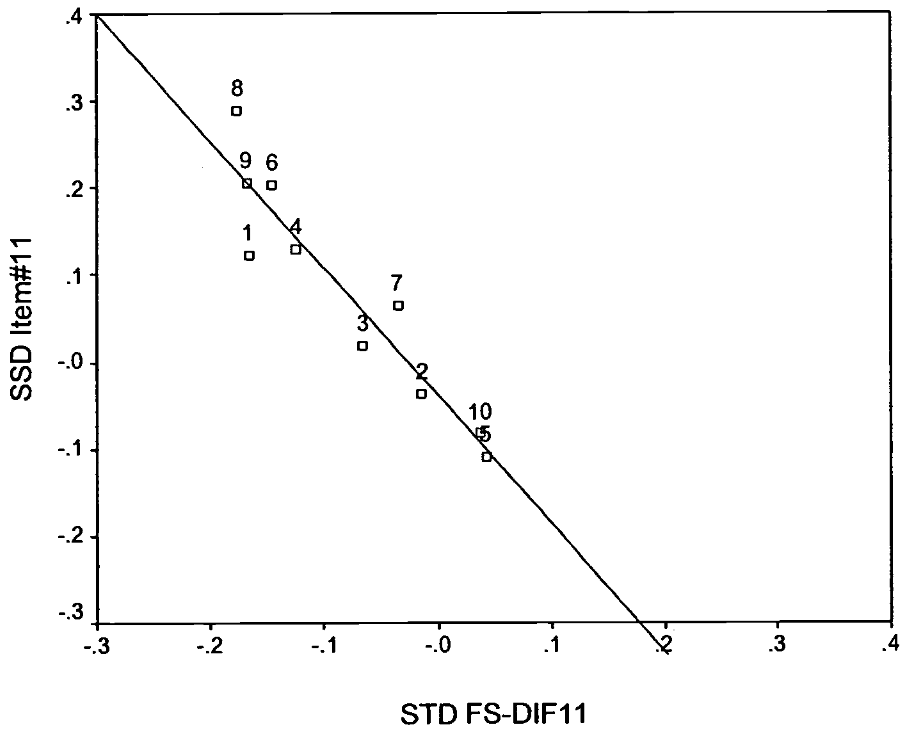
Scatter Plot of Regression Analysis after Removing Item #1



The regression equation is given by $SSD_{Item\ 1} = 0.048 - 1.344 * FS\ STD\ DIF_{Item\ 1}$

Figure 2

Scatter Plot of Regression Analysis after Removing Item #11

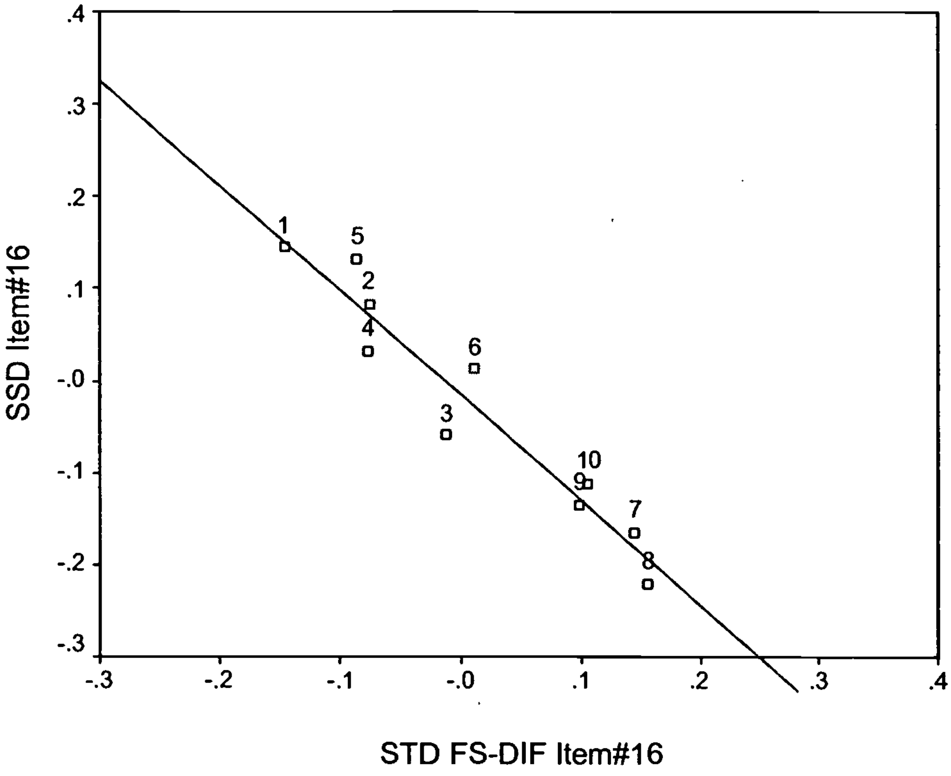


The regression equation is given by

$$SSD_{\text{Item } 11} = -0.040 - 1.465 * \text{FS STD DIF}_{\text{Item } 11}$$

Figure 3

Scatter Plot of Regression Analysis after Removing Item #16



The regression equation is given by

$$SSD_{\text{Item 16}} = -0.016 - 1.139 * \text{FS STD DIF}_{\text{Item 16}}$$

These results show that there is a strong negative relationship between the SSDs and the two-way *FS STD-DIF* indices. Each increase of 0.10 in *FS STD-DIF* index is accompanied by a scaled score reduction of less than one whole scaled score point on the 20 to 80 point scale, a small but noticeable shift nonetheless.

Discussions

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 (the Dissection DIF method). Since different DIF effects exist in each subgroup, it is believed that using a combination of all groups as the reference group produces more accurate, though potentially less stable, findings than using a simple majority group approach.

As we hypothesized, the deletion of DIF items disadvantageous to a particular group has been shown to affect that group the most. Scaled score differences after item deletion and re-equating did vary 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 most when all three items were deleted which was consistent with the fact that they were disadvantaged on all those items. However, the gained and lost points amounted to less than one scaled-point on a 20 to 80 point scale.

We also hypothesized that the effects of item deletion on scores can be predicted based on the standardization method. Regression analyses confirmed that the standardization DIF method can reliably predict score changes. It was shown that by deleting a negative DIF item that the focal group is benefited and by deleting a positive DIF item, the reference group is benefited. However, the sample sizes were not adequately large to generalize the findings.

The purpose of using the dissection classification scheme within the context of a two-way procedure is to examine gender by ethnicity interactions that traditional DIF grouping methods, i.e. one-way methods, do not allow. The dissection classification method places everyone in the reference group simultaneously. In terms of gender by ethnicity interaction, these results show that the among the three item deletion scenarios, the two way DIF effects showed very little interaction except one case: the Asian group. Future work will investigate the nature of this interaction.

The dissection and two-way DIF method may benefit large-scale standardized testing programs. 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. We have 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. In addition, it was shown that by summing weighted two-way DIF values for each ethnic subgroup, one-way DIF indices for ethnic groups can be obtained. The one-way DIF values for gender groups can be derived by entering two-way DIF values of gender subgroups into weighted equations. 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 one-way DIF analysis and two-way DIF method is through the analogy of analysis of variance (ANOVA). 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.

Limitations

A limitation of this study is 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. Another issue worth investigating is to obtain predicted scaled scores on the shortened tests by applying the full test local linear approximation (Dorans, 1984) and then compare them with the observed scaled score values for each focal group.

The substantive DIF findings obtained in this study should be interpreted cautiously. First, the final forms of 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 single pretest, which is less than half the length of actual tests (35 items vs. 78 items). 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.

References

Carlton, S. T., and 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. and Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24*, 157-166.

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. *Journal of Educational Measurement, 23(3)*, 245-264.

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

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

Dorans, N. J., and Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In Bennett, R. E. and 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.

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

Holland, P. W., and Thayer, D. T. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test Validity*, (pp. 129-145). Hillsdale, New Jersey: 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. (1988). Traditional equating methodology. *Educational measurement: Issues and practice. 17 (1),* 29-36.

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

Scheuneman, J. D., and 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., and 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.

ETS Educational Testing Service

Notes

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 I: Verbal 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.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034896

I. DOCUMENT IDENTIFICATION:

Title: Using DIF Dissection to Assess Effects of Item Deletion due to DIF on the Performance of SAT® I: Reasoning Test Subpopulations

Author(s): Yanling Zhang, Joy Matthews-López, Neil J. Dorans

Corporate Source: Educational Testing Service

Publication Date: April, 2003 @ NCME, Chicago

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please →

Signature: Joy M. Lopez —

Printed Name/Position/Title: Joy Matthews-López, Director of Research, Ohio University

Organization/Address:
Rm. 319 Grosvenor Hall
O.U. College of Osteopathic Medicine
Athens, Ohio 45701

Telephone: 740 593-2380

FAX: 740 593-2320

E-Mail Address: joy.matthews-lopet@ohio.edu

Date: 4/20/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:	n/a
Address:	
Price:	

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address: n/a

Name:	
Address:	

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org