. DOCUMENT RESUME

ED 384 659 TM 023 953

AUTHOR	Schmitt, Alicia P.; Crone, Carolyn R.
TITLE	Alternative Mathematical Aptitude Item Types: DIF Issues.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-91-42
PUB DATE	Jul 91
NOTE	39p.; Version of a paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Algebra; Asian Americans; Black Students; *Constructed Response; Criteria; *Difficulty Level; Ethnic Groups; Field Tests; Hispanic Americans; *Item Bias; *Mathematics Tests; Racial Differences; Sex Differences; Student Placement; *Test Items; White Students
IDENTIFIERS	*Alternative Assessment; Mantel Haenszel Procedure; *Scholastic Aptitude Test; Speededness (Tests); Standardization

ABSTRACT

Alternative mathematical items administered as prototypes at the spring 1989 Field Trials of the Scholastic Aptitude Test (SAT) are evaluated for differential item functioning (DIF) and differential speededness. Results for Algebra Placement (AP) and Student Produced Response (SPR) items are presented and contrasted with results obtained on two current SAT-Math (SAT-M) items: Regular Math and Quantitative Comparison. Analyses on comparisons between female and comparable male examinees, and between Asian-American, Black, and Hispanic examinees in comparison to comparable White examinees indicate that both of these alternative items appear to have DIF. Additional DIF analyses comparing the use of an internal versus an external matching criteria for the SPR items show evidence of negative DIF with either criteria. Results using the Mantel Haenszel Delta DIF statistic are more extreme than DIF results using the standardization p metric DIF index. Differential speededness results indicate that the two Math prototypes have slightly higher levels of differential speededness than the SAT-M. The SPR items pose an interesting problem for DIF. The definition of an appropriate DIF matching criterion for constructed response items and metric differences between methods and their effect on difficult or easy items also need further exploration. One figure and 13 tables illustrate the analysis. (Contains 15 references.) (Author/SLD)

RR-91-42

RESEARCH REPORT

ALTERNATIVE MATHEMATICAL APTITUDE ITEM TYPES: DIF ISSUES

Alicia P. Schmitt
Carolyn R. Crone

Educational Testing Service
Princeton, New Jersey
July 1991

Alternative Mathematical Aptitude Item Types: DIF Issues

Alicia P. Schmitt[1,2,3] and Carolyn R. Crone

Educational Testing Service

Running head: ALTERNATIVE MATH ITEMS AND DIF

[1] Earlier version was presented at the National Council on Measurement in Education, Chicago, April, 1991.

[2] The authors thank Liz Burton, Karen Carroll, Neil J. Dorans, and Jeannine Gilsdorf.

[3] Funding for this research was provided by the Admissions Testing Program of the College Board and Educational Testing Service. The interpretation of results represent the opinions of the authors and should not be viewed as statements of ETS or the College Board.

Abstract

Alternative mathematical items administered as prototypes at the Spring 1989 Field Trials are evaluated for differential item functioning (DIF) and differential speededness. Results for Algebra Placement (AP) and Student Produced Response (SPR) items are presented and contrasted to results obtained on the two current SAT-Math items: Regular Math and Quantitative Comparison. Analyses on comparisons between female and comparable male examinees, and between Asian-American, Black, and Hispanic examinees in comparison to comparable White examinees indicate that both of these alternative items appear to have DIF. Additional DIF analyses comparing the use of an internal versus an external matching criteria for the SPR items show evidence of negative DIF with either criteria. Results using the MH D-DIF statistic are more extreme than DIF results using the STD P-DIF index. The metric used to calculate the DIF indices may be accountable for the differences observed. Differential speededness results indicate that the two Math prototypes have slightly higher levels of differential speededness than the SAT-M.

The SPR items pose an interesting problem for DIF. The definition of an appropriate DIF matching criterion for constructed response item types needs more study. Metric differences between methods and their effect on difficult or easy items also needs further exploration. Until these methodological issues are resolved, results of DIF studies on constructed response items should be interpreted with caution.

Alternative Mathematical Ap 'tude Item Types:
DIF Issues

For the past several years the College Board and the Educational Testing Service have engaged in a major endeavor to evaluate possible modifications to the current SAT. As part of this effort, two alternative Math items, Algebra Placement (AP) and Student Produced Response (SPR) items, have been investigated. Investigation of these items have included differential item functioning (DIF). In addition, differential item speededness has also been addressed. For a description of alternatives under consideration for the new SAT I: Math in 1994 see Braswell (1991).

Differential Item Functioning

Differential item functioning refers to a psychometric difference in how an item functions for two comparable groups. Groups are matched with respect to the construct being measured by the test. Comparisons of matched or comparable groups is critical because it is important to distinguish between differences in item functioning and differences in group ability. For descriptions of methods to estimate DIF, refer to: Dorans and Kulick, 1986; Green, Crone, and Folk, 1989; Holland and Thayer, 1988; Scheuneman and Bleistein, 1989; Shepard, Camilli, and Williams, 1985. Examinee response style factors that may be related to DIF deal with how different groups' examinees approach the test taking experience and how they deal with difficult items. One of these factors is differential speededness.

Differential Speededness

When an examinee does not respond to an item and does not respond to any subsequent items in a timed section, all those items are referred to as not reached. It is assumed that all examinees who do not respond to omitted items do so because the items are deemed too difficult. While some examinees may omit not-reached items because they are difficult, it is assumed that most, if not all, examinees who do not respond to a not-reached item, do not, in fact, reach the item. Not-reached ra:es on items at the end of a timed section or test measure the section or test speededness for the group that took the test.

Differential speededness refers to the existence of differential response rates between focal group members and matched reference group members to items appearing at the end of a section. Schmitt and Bleistein (1987) found evidence of this phenomenon for Blacks, as compared to a comparable group of Whites, on analogy items. Schmitt and Dorans (1990) reported this effect for Hispanics as well. Dorans, Schmitt, and Bleistein (1988) examined differential speededness results for Black, Hispanic, and Asian-American focal groups, compared to a White base or reference group. Results from that study indicated higher standardized differential not-reached rates on the last ten items of the two verbal sections of the two SAT-Verbal form studied for Blacks, Mexican-Americans, and Puerto Ricans. No differential not-reached rates were observed for the Asian-American focal group. Schmitt, Dorans, Crone, and Maneckshana (1990) replicated the previous findings for the SAT-Verbal and found that the same pattern was evident for the SAT-Math.

Purpose of Present Study

The purpose of this study is to evaluate differential item functioning and differential speededness for different ethnic and both gender groups on the alternative Math items field tested at the Spring 1989 Field Trials and to contrast results to those obtained for the two current SAT-M items, Regular Mathematics (RM) and Quantitative Comparison (QC). In addition, this study will specifically address the following issues for the Mathematical Student Produced Response items: 1) Whether the computation of DIF using two alternative DIF indices in two different item difficulty metrics affect DIF results; and 2) Whether using an internal versus an external matching criterion affects DIF assessment.

Method

Instrument

Items in one mathematical form of the SAT, and in two mathematical prototypes tested at the Spring 1989 Trials were used in the analyses for this study. The SAT-M test score is composed of 35 five-option RM items and 25 four-option QC items. The SAT-M is administered in two separately-timed 30-minute sections. Whereas Section 1 is comprised entirely of 25 RM items, Section 2 contains both mathematical item types and has the following item order: 7 RM, 25 QC, and 8 RM.

The two mathematical prototypes were composed of regular SAT-M items and of two alternative items: Algebra Placement (AP) and Student Produced Responses (SPR). Each prototype was also administered in two separately-timed 30-minute sections. The composition of their respective sections was:

Prototype Form A:		Prototype Form B:	
Section 1:	15 AP and 25 QC	Section 1:	10 AP and 20 RM
Section 2:	20 SPR	Section 2:	20 QC and 10 SPR.

The configuration of these three tests is presented in Table 1.

Insert Table 1 about here

The AP items were taken from the Descriptive Tests of Mathematics Skills (DTMS) and from regular SAT-M. The SPR items, a constructed response item type, provided students with a grid-in format on the answer sheet which allowed students to grid-in their responses in up to four characters. Each character could represent numerals 0 through 9, a decimal point, or a back-slash (/) to depict fractions. The SPR items were found to be the most difficult item type and to have particularly high item discrimination.

Lehman and Mazzeo (1991) report that the Math prototype Form A is probably not unidimensional, but can be well represented by a model that supports an Algebra subscore (based on all 30 Algebra items) and possibly a second subscore based on the Arithmetic and Geometry items. They found that a 1-factor model probably provides an adequate representation for prototype Form B. Their evaluation of analyses based on parcels formed by item type indicated that SPR items do not introduce an additional item type factor (Lehman & Mazzeo, 1991).

Samples

As part of the Spring 1989 Field Trials, a spiraling plan with 30 subsamples was designed. See Lehman and Mazzeo (1991) for a full description of the spiralling design. Each of the subsamples took two tests. The SAT-M and each Math prototype were administered to seven subsamples each. Of the seven subsamples who took the SAT-M, two also took Math prototype Form A, and another two also took, Math prototype Form B, as their second test. Of the remaining five subsamples who took either Math prototype Form A or B, two received the Math prototypes first. In all five of these subsamples the Math prototypes were administered with a non-math test. Because of the spiraling design, these subsamples were considered random samples and for purposes of the present study, data was aggregated across the seven subsamples. From this pool, members of four ethnic groups, Asian American, Black, Hispanic, and White examinees, and of both genders were selected to create the focal and reference groups used in the DIF analyses. The White group served as the reference group for all other ethnic focal groups, and males as the reference group for the gender comparison. All focal and reference groups used in this study were restricted to college-bound juniors who attended high schools that were not part of the Fall 1988 Field Trials and who identified themselves as having English as one of their first languages. Table 2 presents sample sizes for each of the ethnic and gender groups. This table identifies the samples that were analyzed using an internal criterion (corresponding total score) or an external criterion (total score on the SAT-M -- two subsamples each).

Insert Table 2 about here

Speededness

The speededness of a test is a measure of the degree to which test-takers were able to attempt all test items within the allotted time period. Several indices of speededness were evaluated for the SAT-M form and for the Math prototypes Form A and B based on the performance of the total group. The criterion used at Educational Testing Service to indicate speededness in a test section is that all examinees are able to complete 75% of the section, and that 80% of the examinees are able to complete the entire section. In order to assess whether 80% of the examinees completed the section, two indices are considered: the actual percentage completing the section, and the number of items reached by 80% of the examinees.

The percentage of examinees completing the test may be low because the last item or several items in the test are speeded, or it may be low simply because the last item on the test is difficult. The number of items completed by 80% of the examinees is an index of the number of relatively unspeeded items in the test.

The speededness data for the SAT-M is typical of regularly administered SAT forms. Over 98% of examinees completed 75% of either Section 1 or 2 of the SAT-M. All but one of the items were reached by 80% of the examinees in Section 1 (24 of 25) and all but two were reached by 80% of examinees in Section 2 (33 of 35).

The speededness data for the Math prototype Form A indicates that Section 1 speededness data is typical of regularly administered SAT forms; but that Section 2 (consisting of 20 SPR items) is considerably more speeded. Over 98% of examinees completed 75% of Section 1 while 90% of examinees completed 75% of Section 2 of Form A. All but one of the items were reached by 80% of the examinees in Section 1 (39 of 40) and all but two were reached by 80% of examinees in Section 2 (18 of 20).

The speededness data for the Math prototype Form B shows (as found for Form A) that Section 1 speededness data is typical of regularly administered SAT forms; but that Section 2 (consisting of 20 QC and 10 SPR items) is more speeded. Over 98% of examinees completed 75% of Section 1 while 95.8% of examinees completed 75% of Section 2 of Form B. Eighty percent of the examinees completed 27 of the 30 items in Section 1, and 26 of the 30 items in Section 2.

Form A, Section 2 is the most speeded section. This greater speededness may be a function of the item format, item content, or an interaction of these possible effects.

Statistics

Items that are harder for one group than for another group with the same level of ability or skill are defined as differentially more difficult or as functioning differentially between the two groups. Usually the majority group is referred to as the reference or base group and the minority group as the focal or study group. Since DIF indices take overall differences in ability into account by matching the groups before comparing their item performance, DIF indices identify items that might have construct-irrelevant characteristics. Judgmental evaluation of items with DIF may identify some possible causes of DIF.

Two statistical procedures currently used at the Educational Testing Service to assess DIF are the Mantel-Haenszel (MH) (Holland & Thayer, 1988) and the standardization (STD) methods (Dorans & Kulick, 1986). Both of these methods identify DIF after partitioning the reference and focal groups into subgroups with the same score on a relevant matching variable. The matching variable is usually the total score on a test closely related to the construct that the item is intended to measure. While there are some differences between the MH and STD methods, such as the scale in which the item performance of the reference and focal groups are compared and the way that the individual differences between the subgroups are averaged, when reported on the same metric, the DIF estimates computed by these methods are highly correlated (upper .90's) because they tend to yield the same rank order of items with respect to DIF (Wright, 1987; Holland & Thayer, 1988; Dorans, 1989).

Standardization procedure. In the traditional standardization analysis, an item is said to exhibit differential item functioning when the probability of correctly answering the item is lower or higher for examinees from one group than for equally able examinees from another group. The focus of DIF analyses is on differences in performance between groups that are matched with respect to the ability, knowledge, or skill of interest. The basic elements of a standardization analysis of the keyed response are proportions correct at each level of a matching variable, such as total score, in a base or reference group and a focal or study group. Standardization provides numerical indices for quantifying DIF in the **p metric.**

The prime numerical DIF index that standardization computes is the standardized **p-difference,** which is defined as:

(1) $\text{STD P-DIF} = \Sigma\{W_s[P_{fs} - P_{rs}]\} / \Sigma\{W_s\}$,

where $[W_s / \Sigma\{W_s\}]$ is the weighing factor at score level s on the SAT used to weight differences in the proportions correct between the focal group (P_{fs}) and the reference group (P_{rs}), and Σ is the summation operator which sums these weighted differences across scores levels to arrive at STD P-DIF, an index that can range from -1 to +1 or -100% to 100%. Negative values of STD P-DIF indicate that the item disadvantages the focal group, while positive STD P-DIF values indicate that the item favors the focal group. STD P-DIF values between -.05 (-5%) and +.05 (+5%) are considered negligible. STD P-DIF values outside the {-.10, +.10} or the {-10%, +10%) range are considered sizeable. For operational purposes, a $|\text{STD P-DIF}| \geq .10$ is a recommended cutoff; for research purposes, a cutoff of $|\text{STD P-DIF}| \geq .05$ should be used.

The weights, $[W_s / \Sigma\{W_s\}]$, which are applied to both P_{fs} and P_{rs}, are the essence of the standardization approach. Contrast this weighing constancy with what occurs in the computation of impact,

(2) $\text{IMPACT} = P_f - P_b = \Sigma\{N_{fs}P_{fs}\}/\Sigma\{N_{fs}\} - \Sigma\{N_{rs}P_{rs}\}/\Sigma\{N_{rs}\}$,

where N_{fs} and N_{rs} are the frequencies of score level s in the focal and reference groups. Thus, impact provides differences in performance between groups which have not been matched and are probably not comparable. In contrast, the standardization method focuses on comparing groups that are comparable. In addition, the particular set of weights employed for standardization depends upon the purposes of the investigation. In practice, $W_s = N_{fs}$ has been used because it gives the greatest weight to differences in P_{fs} and P_{rs} at those score levels most often attained by the focal group under study. Use of N_{fs} means that STD P-DIF equals the difference between P_f, the observed performance of the focal group on the item, and P_f, the imputed performance of selected reference group members who are matched in ability to the focal group members.

In research on the SAT, two versions of STD P-DIF have been computed. In the original version, all examinees including those who do not reach the item are included in the denominator of P_{fs} and P_{rs}, yielding STD P-DIF$_1$. In the more recent version, an effort was made to adjust for speededness by excluding the not reached examinees from the calculation of STD P-DIF, yielding STD P-DIF$_2$. Schmitt and Bleistein (1987) were the first to employ this correction, which has become the standard in operational DIF work on the SAT. Dorans, Schmitt, and Curley (1988) demonstrated that this correction partially adjusts for the speededness effect. Hence, STD P-DIF$_2$ will be used in this report for all options except not reached where only STD P-DIF$_1$ makes sense.

The generalization of the standardization methodology to all response options including not reaching the item is straightforward. It is as simple as replacing the keyed response with the option of interest in all calculations. For example, a standardized response rate analysis on not-reached would entail computing the proportions not reaching (NR) (as opposed to the proportions correct (P)) in both the focal and reference groups,

(3) $P_{fs}(NR) = NR_{fs}/N_{fs}$; $NR_{rs}(NR) = NR_{rs}/N_{rs}$,

where NR_{fs} and NR_{rs} are the number of people in the focal and reference groups, respectively, at score level s who did not reach. The next step is to compute differences between these proportions,

(4) $D_s(NR) = P_{fs}(NR) - P_{rs}(NR)$.

Then these individual score level differences are summarized across score levels by applying some standardized weighing function to these differences to obtain STD P-DIF(NR),

(5) $\text{STD P-DIF}(NR) = \Sigma\{W_s[P_{fs}(NR) - P_{rs}(NR)]\} / \Sigma\{W_s\}$,

the standardized difference in not-reached rates. In a similar fashion one can compute standardized differences in response rates for options A, B, C, D, and E, and for omits as well.

For items at the end of a separately-timed section of a test such as the SAT, these standardized differences provide measurement of the differential speededness of a test. Differential speededness refers to the existence of differential response rates between focal group members and matched reference group members to items appearing at the end of a section.

Mantel-Haenszel method. The Mantel-Haenszel (Mantel & Haenszel, 1959) procedure, adapted by Holland and Thayer (1988) for DIF analysis computes ratios of the conditional odds of successful reference group performance over the conditional odds of successful focal group performance at each score level, and then averages these ratios across score levels.

(6) $\alpha_s = [R_{rs}/W_{rs}]/[R_{fs}/W_{fs}] = [R_{rs}W_{fs}]/[R_{fs}W_{rs}]$,

where: (R_{rs}) is the proportion correct for the reference group; (W_{fs}) is the proportion incorrect for the focal group; (R_{fs}) is the proportion correct for the focal group; and (W_{rs}) is the proportion incorrect for the reference group. In the calculation of the average ratio, statistically optimal weights are used for each ratio. The Mantel-Haenszel method provides an estimate of the constant odds-ratio, which is defined as:

(7) $\alpha MH = [\Sigma_s R_{rs} W_{fs}/N_{ts}]/[\Sigma_\sigma R_{fs} W_{rs}/N_{ts}]$.

The Mantel-Haenszel statistic is transformed to the **"delta" metric** used in the ETS test development process. The **"delta" metric** has a mean of 13 and a standard deviation of 4. To obtain a delta, the proportion correct (p) is converted to a z-score via a p-to-z transformation using the inverse of the normal cumulative function, followed by a linear transformation to a metric with a mean of 13 and a standard deviation of 4 via:

(8) $\Delta = 13 - 4\ \{\Phi^{-1}(p)\}$,

such that large values of Δ correspond to difficult items, while easy items have small values of delta. Holland and Thayer (1988) converted αMH into a difference in deltas via:

(9) MH D-DIF= $-2.35 \ln[\alpha\text{MH}]$.

This estimate provides an estimate of DIF effect size on the delta metric. It ranges from negative ∞ to infinity with a value of 0 indicating no DIF. MH D-DIF values between -1.00 and +1.00 are considered negligible. MH D-DIF values outside the {-1.50, +1.50} range are considered sizeable. For operational purposes, a |MH D-DIF|≥1.50 is a recommended cutoff; a |MH D-DIF|≥1.00 but less than 1.50 should be examined for research purposes. Note that positive values of MH D-DIF favor the focal group, while negative values favor the reference group. For a complete description and comparison of the STD P-DIF and MH D-DIF statistics refer to Dorans and Holland (1989).

Procedure

The standardization method was used to compute differential speededness for items on the SAT-M, prototype Form A, and prototype Form B test forms. Differential item functioning statistics were calculated using the Mantel-Haenszel method in the **"delta" metric** and the standardization method in the **p metric**. Because these two DIF computation methods have been found to be closely related when computed in the same metric, any differences were expected to reflect differences due to metric used.

Internal Matching Criteria. The analysis of differential item functioning involved a two-step process to refine the matching criteria. During the first step the matching criterion for the DIF analyses was the total-test raw score on either the SAT-M or each of the Math prototypes (Forms A or B), henceforth referred to as the total score internal matching criterion. On the basis of the initial analysis, any item with extreme DIF values for the corresponding focal group comparison was removed as part of the total score used to match the reference and focal groups. Thus, a "refined" matching criterion was determined for each focal group comparison and each test. Two refined scores were required for each test. Accordingly, the refined internal matching criteria were as follows: 1) For the SAT-M: White/Asian-American: SAT-M = 58 (60 items - 2 items); White/ Hispanics: SAT-M = 59 (60 items - 1 item). 2) For the Math prototype Form A: Male/Female: Form A = 59 (60 items -1 item); White/Black: Form A = 57 (60 items - 3 items). 3) For the Math prototype Form B: Male/Female: Form A = 59 (60 items -1 item); White/Black: Form A = 59 (60 items - 1 item).

Using the refined internal criteria, a second set of analyses were computed. In this second analysis, new DIF statistics were obtained for those items in all reference group/focal group comparisons that were part of the refined internal criterion. DIF values from the initial analysis were used for those items that were not part of the refined internal criteria. Because of the indications of speededness observed for the two Math prototypes, their use as internal matching criteria might affect the appropriateness of the matching, and consequently, the results of the DIF analyses.

External Matching Criteria. For the two sets of Form A and Form B subsamples which had also taken the SAT-M form, additional DIF analyses for the Student Produced Response item type were computed using the SAT-M as external matching criterion. The respective refined SAT-M criteria were used for the White/Asian-American and the White/Hispanic comparisons. The total SAT-M score was used as criteria for the other comparisons.

Results and Discussion

Results of DIF analyses using internal matching criteria are presented first. Extended DIF analyses of the SPR item type using the external SAT-M matching criterion and compared to the DIF findings using internal criteria are discussed second. Differential speededness findings are presented third.

DIF and differential speededness analyses are summarized using descriptive statistics and by a graphical display of speededness patterns of Male/Female, White/Black, White/Hispanics, and White/Asian-American comparisons. All DIF analyses are reported by Mathematical Form, item type, and DIF method-metric.

Differential Item Functioning

Internal Matching Criteria.

Tables 3 and 4 present DIF summary statistics for the SAT-M in the MH (delta-metric) and standardization (p-metric) methods, respectively. Because, in most cases, item types are represented by a limited number of items, the median and percentiles will provide more stable indicators of DIF by item type.

Insert Tables 3 and 4 about here

Differential item functioning computed via the MH D-DIF statistic indicate negligible DIF for all four focal groups for both the RM and the QC item types of the SAT-M. Similar findings were observed when DIF was computed using the STD P-DIF index. Here again, negligible indications of DIF were observed for the RM and the QC items of the SAT-M.

Tables 5 and 6 display the MH D-DIF and the STD P-DIF summary statistics, respectively, for the Math prototype, Form A item types. The MH D-DIF summary statistics indicate that the AP item type has positive DIF across all focal groups, the QC a slight trend toward negative DIF for all focal groups, while the SPR item type has negative DIF, particularly for the Female and Black focal groups. Similar direction of results were observed when DIF was computed using the STD P-DIF index; but the magnitude of the DIF was much smaller, beyond that expected given the relationship reported by Wright (1987).

Insert Tables 5 and 6 about here

The Math prototype, Form B item types' MH D-DIF and STD P-DIF summary statistics are presented in Tables 7 and 8, respectively. Both the MH D-DIF and the STD P-DIF summary statistics indicate that, as observed for Form A, the AP item type has positive DIF for all focal groups, the QC a very slight trend toward negative DIF for all focal groups, and negative DIF for the SPR item type is evident primarily for the Female and Black focal groups. The other item type of Form B, RM, has negligible DIF. Results obtained using the MH D-DIF index were more extreme than results obtained with the STD P-DIF statistic.

Insert Tables 7 and 8 about here

The two alternative item types under evaluation, AP and SPR, tend to exhibit more DIF than typically seen on SAT-M items; but the magnitude of the DIF is more extreme when the MH D-DIF index was used. The AP item type appears differentially easier for Females, Blacks, Hispanics, and Asian-American focal groups.

The SPR item type appears differentially harder for Female and Black focal groups. Nevertheless, there are psychometric and educational advantages to a constructed response math item. Several advantages of the SPR items as seen by the Mathematical education community have been addressed by Braswell (1991). According to Braswell (1990), the SPR item type provides a more natural problem-solving task where the focus is problem-solving skills rather than test-taking skills. DIF results for the SPR items were evaluated further. Because the matching criteria used for Form A and B were their corresponding total scores (internal matching criteria) which were partly composed of a very easy item type, AP, and possibly flawed because of speededness, the DIF results for the SPR could be questionable. Reanalysis using an external matching criterion, which did not include AP or SPR items, were performed by Dorans and Schmitt (1990). These results are discussed in the next section.

External Matching Criteria.

The use of the SAT-M as external matching criterion for DIF analyses of the SPR items and comparison to the internal matching criterion results are presented in this section. Note that because only two subsamples took both SAT-M and Math Form A or B, sample sizes for all focal groups are considerably reduced (see Table 2). A minimum sample size of 200 was specified for focal or reference groups. Sample sizes were insufficient for the Hispanic focal group on Form A and B and for the Asian-American focal group on Form B.

Differential item functioning results via the MH D-DIF and the STD P-DIF methods for the 20 SPR items of Form A are presented in Tables 9 and 10, respectively. The top part of the tables display the SPR DIF results obtained using the internal criterion, while results on the bottom portion present SPR DIF results after matching with the SAT-M external criterion. The SPR items were categorized into eight groupings according to their MH D-DIF or STD P-DIF values. The top four groupings represent positive DIF values where the focal group did differentially better than its reference group, and the bottom four groupings represent negative DIF where the focal group did differentially worse than its reference group.

Insert Tables 9 and 10 about here

Although all focal groups demonstrated negative DIF, Females and Black examinees had more extreme negative DIF items and larger DIF means. Reanalyses using the SAT-M as an external criterion, which did not include AP or SPR items, did not reduce the magnitude or direction of the DIF found using the internal matching criterion. The external matching criterion seemed to increase the negative DIF found for the Black focal group. For the Black focal group the SPR items (Form A) DIF means increased in magnitude from a -.75 for the MH D-DIF with the internal criterion to -.88 with the external criterion and from -.03 for the STD P-DIF with the internal criterion to -.04 with the external criterion. The percentage of negative items for this focal group also increased. Results for the Female focal group appear to have been only slightly reduced when the external criterion was used: the MH D-DIF mean of -.52 with the internal criterion was reduced to -.46 with the external criterion (the STD P-DIF means did not change) and the percentage of negative items was slightly reduced.

Parallel DIF results for the 10 SPR items of Form B using the MH D-DIF and the STD P-DIF methods are presented in Tables 11 and 12, respectively.

Insert Tables 11 and 12 about here

Consistent with the findings for Form A, estimations of DIF after matching with an internal criterion indicate that although all focal groups exhibit negative DIF, Female and Black examinees had more extreme negative DIF items and larger DIF means. Reanalyses using the SAT-M as an external criterion, for the Male/Female and White/Black comparisons, seemed to increase the negative DIF found for both the Female and the Black focal groups. For the Black focal group, the SPR items (Form B) DIF means increased in magnitude from a -.85 for the MH D-DIF with the internal criterion to -.88 with the external criterion and from -.03 for the STD P-DIF with the internal criterion to -.04 with the external criterion. The percentage of negative items for this focal group also increased. While the DIF means for the Female focal group were slightly reduced when the external criterion was used, the percentage of negative items was increased.

Results for the DIF analyses using the external all multiple-choice SAT-M matching criterion did not reduce the DIF found using the internal matching criterion. The external matching criterion seemed to increase the negative DIF found for the Black focal group on the SPR items of both Forms A and B. Although results for the Male/Female comparison were not as consistently more negative across both forms, they were either basically the same (Form A) or much worse (Form B). Consistent with the comparisons between method-metric presented in the previous section, the MH D-DIF statistic rendered more extreme DIF. These differences between the DIF estimation methods may well be a result of their metric. Wright (1987) indicates that the STD P-DIF index is more stable. He accounts for the greater stability of the STD P-DIF index: "The D_{STD} (STD P-DIF) index may be inherently more stable than the odds ratio statistics (MH D-DIF) because it is bounded at the extremes" (p. 10). For this same reason the STD P-DIF statistic may be less affected by very easy or very hard items, such as the harder SPR items.

Differential Speededness

Results of the differential speededness effect found for the Male/Female, White/Black, White/Hispanic, and White/Asian-American comparisons are presented in Table 13 and displayed in Figure 1 for the last ten items of the two sections of the SAT-M, and of the two Math prototypes: Forms A and B.

Insert Table 13 and Figure 1 about here

Examination of the not-reached standardized differences for the last ten items of each SAT-M section show that Black, Hispanic, and Asian-American focal groups have a slightly higher not-reached proportion than the matched White reference group for the last three items of Section 1 and for the last five items of Section 2. While in Section 1 of the SAT-M there does not appear to be much differentiation or differential speededness found for the ethnic focal groups, in Section 2 (the more speeded 35-item section) the Hispanic focal group has the highest proportion of differential not-reached, followed closely by the Asian-American and Black focal groups. No indication of differential speededness was observed, on either section, for the Female focal group.

Results for the last ten items of each Math prototype Form A section show minimal differential not-reached proportions for the Black and Hispanic focal groups and no differential not-reached differences for the Asian-American or Female focal groups. On Section 2 of Form A, more differential speededness is observed for the Black and Hispanic focal groups, with the Black focal group exhibiting greater differential speededness. This is the Section with 20 SPR items at the end. No differential not-reached differences were noticeable for the Asian-American or Female focal groups.

Results for Math prototype Form B, Section 1 has slight differential speededness for the last items for all ethnic subgroups. Hispanic examinees show the larger differential not-reached rates, followed by Blacks and Asian-Americans. Results for Section 2 of Form B are similar to the speededness pattern observed on Section 1. The Hispanic focal group has the highest proportion of differential not-reached, followed closely by the Black focal group. The Asian-American focal group has a light increment in differential speededness for the last four items while the Female focal group shows a decrease in not-reached rates for the last items. As with Section 2 of Form A, the last ten items of Section 2 of Form B is also composed of SPR items.

Conclusions

The present study examined DIF and differential non-responses for the alternative Math items tested at the Spring 1989 Field Trials: AP and SPR and contrasted results to those obtained for the two current SAT-M items: RM and QC. This study also addressed 1) whether computations of DIF with the MH D-DIF or the STD P-DIF methods-metrics affected results and 2) whether the use of an internal versus an external criterion affected DIF results for the SPR items.

The two alternative item types under evaluation, AP and SPR, exhibit more DIF than typically seen on SAT-M items. The AP item type appears differentially easier for Females, Blacks, Hispanics, and Asian-American focal groups. The SPR item type appears differentially harder for Female and Black focal groups. Nevertheless, there are psychometric and educational advantages to a constructed response math item. Several advantages of the SPR items as seen by the Mathematical education community have been addressed by Braswell (1991). Presently, the plan is to have the SPR items follow the same procedures for assembly that all SAT multiple-choice items now do. They will undergo regular item and DIF screening before a test is assembled for final administration. The specifics of the DIF screening procedures, however, may have to be modified in order to reflect DIF issues particular to constructed response item types (e.g., appropriate matching criterion and interaction between DIF method-metric and item difficulty). These issues involve unsolved methodological problems about how to analyze SPR items for DIF.

Additional DIF analyses for the SPR items using an external all multiple-choice SAT-M matching criterion did not reduce the DIF found using the internal matching criterion. The external matching criterion seemed to increase the negative DIF found for the Black focal group on the SPR items of both Forms A and B. Although results for the Male/Female comparison were not as consistently more negative across both forms, they were either basically the same (Form A) or much worse (Form B). Results using the MH D-DIF statistic are more extreme than DIF results using the STD P-DIF index. The metric used to calculate the DIF indices may be accountable for the differences observed. Kulick and Hu (1989) reported high correlations between item difficulty and DIF. Because the alternative items are either very easy, AP, or very hard, SPR, the difficulty of these items seems to be interacting with the metric of the DIF indices used. The STD P-DIF, in the p-metric, is bounded at the extremes; while the MH D-DIF index, in the delta-metric, is not.

Differential speededness results indicate that the two Math prototypes had a slightly higher level of differential speededness than the SAT-M. The pattern of differential speededness by focal groups is consistent with findings for the SAT-Verbal (Schmitt & Dorans, 1990) and for the SAT-M (Schmitt et al., 1990).

The results obtained for the SPR items question the appropriateness of using either the total math internal or external matching criterion. The total math scores are mainly or completely composed of multiple-choice items which are corrected for guessing. The SPR items are a constructed response type of item with no correction for guessing. Constructed response item types may present a very different mode of evaluating knowledge than do multiple-choice formats. If the evaluation of knowledge is affected by the nature of the evaluation task (as may be the case with multiple-choice and constructed response items), then a matching criterion based mainly on items in one format may not be appropriate for DIF estimations of items in the other format. Dorans and Schmitt (1990) note when addressing the use of a multiple-choice matching criterion for constructed response items: "Matching criterion appropriateness is essential for proper DIF analyses of any item type. Finding an appropriate matching criterion for constructed response items may be problematic " (p.31).

The SPR items pose an interesting problem for DIF. The definition of an appropriate DIF matching criterion for constructed response item types needs more study. Metric differences between methods and their effect on DIF assessment for difficult or easy items also needs further exploration. Until these methodological issues are resolved, results of DIF studies on constructed response items should be interpreted with caution.

References

Braswell, J. (1991, April). *Overview of changes in the SAT Mathematics Test in 1994.* Paper presented at the National Council on Measurement in Education, Chicago, IL.

Dorans, N. J., & Holland, P. W. (1989, October). *DIF detection and description: Mantel-Haenszel and standardization.* Paper presented at the ETS Conference, Differential Item Functioning: Theory and Practice, Princeton, NJ.

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

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1988). *The standardization approach to assessing differential speededness* (RR-88-31). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Schmitt, A. P., & Curley, W. E. (1988, March). *Differential speededness: Some items have DIF because of where they are, not what they are.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Green, B. F., Crone, C. R., and Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement, 26,* 147-160.

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

Lehman, J. D., & Mazzeo, J. (1991, April). *Confirmatory factor analyses of Mathematical prototypes.* Paper presented at the National Council on Measurement in Education, Chicago, IL.

Kulick, E., & Hu, P. G. (1989). Examining the relationship between differential item functioning and item difficulty (CB-89-5). New York: College Entrance Examination Board.

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education,* 2, 255-275.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement, 27,* 1-13.

Schmitt. A. P., Dorans, N. J., Crone, C. R., & Maneckshana, B. T. (1990, April). *Differential item omit and speededness patterns on the SAT.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, MA.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22,* 77-105.

Wright, D. J. (1987). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test* (RM-87-1). Princeton, NJ: Educational Testing Service.

Table 1

Configuration of Mathematical Tests
by Section, Item Type, and Item Position
Spring Trials 1989 Administration

Section	Item Type	Item Position
	SAT-Math (60 items)	
Section 1 (25 items)	25 - Regular Math (5-ch)	1-25
Section 2 (35 items)	15 - Regular Math (5-ch)	1-7, 28-35
	20 - Quant. Comp. (4-ch)	8-27
	Math Form A (60 items)	
Section 1 (40 items)	15 - Algebra Placement (4-ch)	1-15
	25 - SAT Quant. Comp. (4-ch)	16-40
Section 2 (20 items)	20 - Student Produced Resp.	1-20
	Math Form B (60 items)	
Section 1 (30 items)	10 - Algebra Placement (4-ch)	1-10
	20 - SAT Regular Math (5-ch)	11-30
Section 2 (30 items)	20 - SAT Quant. Comp. (4-ch)	1-20
	10 - Student Produced Resp.	21-30

Table 2

Sample Sizes of Groups in Math DIF Analyses
For **Internal or External Criteria**
Spring Trials 1989 Administration

	Groups					
Tests	Male	Female	White	Black	Hispanic	Asian A
			Internal Criterion			
SAT-M	5,974	7,249	9,909	1,829	614	688
Math A	6,088	7,129	9,943	1,742	641	728
Math B	6,125	7,490	10,112	1,900	661	749
			External Criterion			
Math A	1,655	1,992	2,717	527	163	202
Math B	1,650	2,001	2,699	533	178	190

Note. Samples for DIF analyses were restricted to college-bound juniors who did not attend high schools that were part of the Fall 1988 Field Trials and who identified themselves as having English as one of their first languages. Two Math prototypes were administered at the Spring 1989 Trials: Math Forms A and B.

Table 3

Summary Statistics of MH D-DIF
for SAT-Math by Item Type
Spring Trials 1989 Administration

	Comparison			
SUMMARY STATISTICS[1]	MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISP.	WHITE/ ASIAN A.
SAT-M (40 Regular Math items)				
Mean	-0.07	-0.07	-0.08	-0.11
S.D.	0.46	0.42	0.57	0.66
Maximum	0.73	0.70	0.79	1.36
90%-Tile	0.53	0.42	0.57	0.69
Median	-0.00	-0.05	0.02	0.01
10%-Tile	-0.78	-0.59	-0.75	-0.81
Minimum	-1.30	-1.11	-1.90	-2.03
SAT-M (20 Quantitative Comparison items)				
Mean	0.01	0.06	-0.01	-0.01
S.D.	0.43	0.43	0.27	0.40
Maximum	0.69	0.82	0.57	0.82
90%-Tile	0.57	0.58	0.26	0.55
Median	0.01	0.13	0.08	-0.02
10%-Tile	-0.65	-0.57	-0.38	-0.58
Minimum	-0.82	-1.10	-0.61	-0.65

[1] Internal matching criterion was used.

Table 4

Summary Statistics of **STD P-DIF** for **SAT-Math** by Item Type Spring Trials 1989 Administration

	Comparison			
SUMMARY STATISTICS[1]	MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISP.	WHITE/ ASIAN A.
	SAT-M (40 Regular Math items)			
Mean	-0.00	-0.00	-0.01	-0.01
S.D.	0.03	0.03	0.04	0.04
Maximum	0.04	0.05	0.06	0.06
90%-Tile	0.03	0.03	0.04	0.03
Median	-0.00	-0.00	0.00	0.00
10%-Tile	-0.05	-0.04	-0.05	-0.05
Minimum	-0.08	-0.09	-0.14	-0.15
	SAT-M (20 Quantitative Comparison items)			
Mean	0.00	0.01	0.00	0.00
S.D.	0.03	0.04	0.02	0.03
Maximum	0.05	0.09	0.05	0.06
90%-Tile	0.04	0.04	0.02	0.04
Median	0.00	0.01	0.01	-0.00
10%-Tile	-0.05	-0.05	-0.03	-0.04
Minimum	-0.06	-0.08	-0.04	-0.04

[1] Internal matching criterion was used.

Table 5

Summary Statistics of MH D-DIF
for Prototype **Math Form A** by Item Type
Spring Trials 1989 Administration

	Comparison			
SUMMARY STATISTICS[1]	MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISP.	WHITE/ ASIAN A.
Prototype Math Form A (15 Algebra Placement items)				
Mean	0.63	0.58	0.55	0.56
S.D.	0.26	0.42	0.40	0.42
Maximum	1.08	1.73	1.46	1.09
90%-Tile	0.92	1.26	0.88	1.06
Median	0.71	0.49	0.68	0.71
10%-Tile	0.17	0.11	-0.10	-0.24
Minimum	0.14	0.00	-0.14	-0.27
Prototype Math Form A (25 SAT-Quantitative Comparison items)				
Mean	-0.17	-0.06	-0.19	-0.15
S.D.	0.44	0.36	0.33	0.43
Maximum	0.78	0.82	0.54	0.82
90%-Tile	0.43	0.27	0.30	0.47
Median	-0.14	-0.01	-0.28	-0.24
10%-Tile	-0.65	-0.47	-0.68	-0.56
Minimum	-1.08	-0.90	-0.74	-1.09
Prototype Math Form A (20 Student Produced Response items)				
Mean	-0.52	-0.75	-0.18	-0.18
S.D.	0.64	0.77	0.53	0.59
Maximum	0.46	0.44	1.05	0.73
90%-Tile	0.27	0.11	0.46	0.69
Median	-0.49	-0.65	-0.19	-0.11
10%-Tile	-1.42	-1.94	-0.67	-0.96
Minimum	-1.86	-2.52	-1.51	-1.02

[1] Internal matching criterion was used.

Table 6

Summary Statistics of **STD P-DIF**
for Prototype **Math Form A** by Item Type
Spring Trials 1989 Administration

	Comparison			
SUMMARY STATISTICS[1]	MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISP.	WHITE/ ASIAN A.
Prototype Math Form A (15 Algebra Placement items)				
Mean	0.04	0.04	0.04	0.03
S.D.	0.02	0.03	0.03	0.02
Maximum	0.08	0.14	0.11	0.08
90%-Tile	0.07	0.10	0.07	0.07
Median	0.04	0.04	0.04	0.03
10%-Tile	0.01	0.01	-0.01	-0.01
Minimum	0.01	0.00	-0.01	-0.01
Prototype Math Form A (25 SAT-Quantitative Comparison items)				
Mean	-0.01	0.00	-0.01	-0.01
S.D.	0.03	0.03	0.02	0.03
Maximum	0.06	0.05	0.04	0.05
90%-Tile	0.02	0.03	0.02	0.03
Median	-0.01	-0.00	-0.02	-0.02
10%-Tile	-0.06	-0.05	-0.04	-0.04
Minimum	-0.07	-0.07	-0.06	-0.06
Prototype Math Form A (20 Student Produced Response items)				
Mean	-0.03	-0.03	-0.01	-0.02
S.D.	0.05	0.04	0.02	0.04
Maximum	0.03	0.02	0.05	0.05
90%-Tile	0.02	0.01	0.02	0.03
Median	-0.03	-0.02	-0.01	-0.01
10%-Tile	-0.11	-0.10	-0.04	-0.06
Minimum	-0.12	-0.12	-0.06	-0.09

[1] Internal matching criterion was used.

Table 7

Summary Statistics of **MH D-DIF**
for Prototype **Math Form B** by Item Type
Spring Trials 1989 Administration

	Comparison			
SUMMARY STATISTICS[1]	MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISP.	WHITE/ ASIAN A.
Prototype Math Form B (10 Algebra Placement items)				
Mean	0.67	0.55	0.58	0.73
S.D.	0.23	0.37	0.37	0.37
Maximum	1.06	1.39	1.16	1.28
90%-Tile	0.99	1.08	1.14	1.25
Median	0.68	0.55	0.61	0.72
10%-Tile	0.33	0.06	0.06	0.18
Minimum	0.26	-0.23	0.05	-0.02
Prototype Math Form B (20 SAT Regular Math items)				
Mean	-0.07	0.07	0.04	0.14
S.D.	0.63	0.53	0.35	0.49
Maximum	1.05	0.85	0.58	0.90
90%-Tile	0.76	0.66	0.54	0.80
Median	0.03	0.17	0.13	0.17
10%-Tile	-0.91	-0.81	-0.49	-0.52
Minimum	-1.34	-1.15	-0.65	-0.72
Prototype Math Form B (20 SAT Quantitative Comparison items)				
Mean	-0.14	-0.21	-0.18	-0.18
S.D.	0.39	0.48	0.51	0.42
Maximum	0.58	1.08	1.18	1.18
90%-Tile	0.38	0.35	0.44	0.15
Median	-0.06	-0.27	-0.30	-0.19
10%-Tile	-0.70	-0.85	-0.80	-0.65
Minimum	-0.85	-0.92	-0.99	-0.93
Prototype Math Form B (10 Student Produced Response items)				
Mean	-0.69	-0.85	-0.23	-0.39
S.D.	0.57	0.64	0.69	0.57
Maximum	0.24	0.16	1.73	1.08
90%-Tile	0.14	0.12	0.85	0.47
Median	-0.67	-1.00	-0.48	-0.43
10%-Tile	-1.49	-1.68	-0.72	-0.98
Minimum	-1.88	-2.00	-0.76	-1.02

[1] Internal matching criterion was used.

Table 8

Summary Statistics of STD P-DIF
for Prototype **Math Form** B by Item Type
Spring Trials 1989 Administration

	Comparison			
SUMMARY STATISTICS[1]	MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISP.	WHITE/ ASIAN A.
Prototype Math Form B (10 Algebra Placement items)				
Mean	0.04	0.04	0.04	0.04
S.D.	0.02	0.03	0.03	0.02
Maximum	0.07	0.12	0.09	0.07
90%-Tile	0.07	0.09	0.09	0.07
Median	0.04	0.04	0.04	0.04
10%-Tile	0.02	0.00	0.00	0.01
Minimum	0.01	-0.02	0.00	0.00
Prototype Math Form B (20 SAT Regular Math items)				
Mean	-0.01	0.00	0.00	0.00
S.D.	0.04	0.03	0.02	0.03
Maximum	0.08	0.04	0.04	0.05
90%-Tile	0.04	0.04	0.04	0.04
Median	0.00	0.01	0.01	0.01
10%-Tile	-0.06	-0.04	-0.03	-0.04
Minimum	-0.12	-0.08	-0.04	-0.05
Prototype Math Form B (20 SAT Quantitative Comparison items)				
Mean	-0.01	-0.01	-0.02	-0.01
S.D.	0.03	0.04	0.03	0.02
Maximum	0.04	0.07	0.04	0.03
90%-Tile	0.02	0.04	0.02	0.01
Median	-0.00	-0.01	-0.02	-0.01
10%-Tile	-0.05	-0.06	-0.05	-0.04
Minimum	-0.05	-0.08	-0.08	-0.06
Prototype Math Form B (10 Student Produced Response items)				
Mean	-0.05	-0.03	-0.02	-0.03
S.D.	0.04	0.03	0.02	0.04
Maximum	0.00	0.01	0.01	0.07
90%-Tile	0.00	0.01	0.01	0.03
Median	-0.04	-0.02	-0.02	-0.03
10%-Tile	-0.10	-0.09	-0.05	-0.07
Minimum	-0.12	-0.11	-0.05	-0.07

[1] Internal matching criterion was used.

Table 9

Differential Item Functioning (DIF) Summary
For MH D-DIF on SPR Items of
Math Form A Using Internal or External Criteria
Spring Trials 1989 Administration

		Category of DIF Value For All Comparisons				
	CROSS-GROUP[a]	CROSS-GROUP[a]	MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISPANIC	WHITE/ ASIAN
MH D-DIF Category	Number	% of Items	Percent of Items by DIF Category			
		Internal Criterion				
DIF ≥ 1.5	0	0.0	0.0	0.0	0.0	0.0
1.0≤ DIF < 1.5	1	5.0	0.0	0.0	5.0	0.0
0.5≤ DIF < 1.0	1	5.0	0.0	0.0	5.0	15.0
0.0≤ DIF < 0.5	4	20.0	25.0	15.0	35.0	25.0
-0.5< DIF < 0.0	1	5.0	25.0	25.0	30.0	20.0
-1.0< DIF ≤-0.5	3	15.0	30.0	25.0	20.0	35.0
-1.5< DIF ≤-1.0	5	25.0	10.0	15.0	0.0	5.0
DIF ≤-1.5	5	25.0	10.0	20.0	5.0	0.0
Total	20	100.0	100.0	100.0	100.0	100.0
Mean	--	---	-0.52	-0.75	-0.18	-0.18
S.D.	--	---	0.64	0.77	0.53	0.59
Maximum	--	---	0.46	0.44	1.05	0.73
Minimum	--	---	-1.86	-2.52	-1.51	-1.02
		External Criterion				
DIF ≥ 1.5	0	0.0	0.0	0.0	N/A	0.0
1.0≤ DIF < 1.5	3	15.0	0.0	0.0	N/A	15.0
0.5≤ DIF < 1.0	3	15.0	10.0	0.0	N/A	20.0
0.0≤ DIF < 0.5	3	10.0	20.0	5.0	N/A	20.0
-0.5< DIF < 0.0	0	0.0	20.0	35.0	N/A	10.0
-1.0< DIF ≤-0.5	4	20.0	30.0	30.0	N/A	30.0
-1.5< DIF ≤-1.0	2	10.0	15.0	10.0	N/A	5.0
DIF ≤-1.5	5	25.0	5.0	20.0	N/A	0.0
Total	20	100.0	100.0	100.0	N/A	100.0
Mean	--	---	-0.46	-0.88	N/A	0.04
S.D.	--	---	0.71	0.91	N/A	0.84
Maximum	--	---	0.82	0.12	N/A	1.26
Minimum	--	---	-1.78	-3.76	N/A	-1.27

[a] Each item is identified in only one DIF category. If the item was flagged for more than one comparison analysis then the largest absolute DIF value indicates its category across all comparisons.

[b] N/A - Insufficient sample size (N < 200) for DIF analysis.

Table 10

Differential Item Functioning (DIF) Summary
For **STD P-DIF** on **SPR** Items of
Math Form A Using **Internal** or **External Criteria**
Spring Trials 1989 Administration

		Category of DIF Value For All Comparisons				
	CROSS-GROUP[a]	CROSS-GROUP[a]	MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISPANIC	WHITE/ ASIAN
STD P-DIF Category	Number	% of Items	Percent of Items by DIF Category			
		Internal Criterion				
DIF ≥ .15	0	0.0	0.0	0.0	0.0	0.0
.10≤ DIF < .15	0	0.0	0.0	0.0	0.0	0.0
.05≤ DIF < .10	0	0.0	0.0	0.0	0.0	0.0
.00≤ DIF < .05	7	35.0	25.0	15.0	40.0	40.0
-.05< DIF < .00	4	20.0	45.0	60.0	55.0	20.0
-.10< DIF ≤-.05	4	20.0	15.0	15.0	5.0	40.0
-.15< DIF ≤-.10	5	25.0	15.0	10.0	0.0	0.0
DIF ≤-.15	0	0.0	0.0	0.0	0.0	0.0
Total	20	100.0	100.0	100.0	100.0	100.0
Mean	--	---	-0.03	-0.03	-0.01	-0.02
S.D.	--	---	0.05	0.04	0.02	0.04
Maximum	--	---	0.03	0.02	0.05	0.05
Minimum	--	---	-0.12	-0.12	-0.06	-0.09
		External Criterion				
DIF ≥ .15	0	0.0	0.0	0.0	N/A	0.0
.10≤ DIF < .15	0	0.0	0.0	0.0	N/A	0.0
.05≤ DIF < .10	5	25.0	5.0	0.0	N/A	20.0
.00≤ DIF < .05	4	20.0	25.0	5.0	N/A	35.0
-.05< DIF < .00	3	15.0	35.0	60.0	N/A	15.0
-.10< DIF ≤-.05	4	20.0	20.0	25.0	N/A	30.0
-.15< DIF ≤-.10	4	20.0	15.0	10.0	N/A	0.0
DIF ≤-.15	0	0.0	0.0	0.0	N/A	0.0
Total	20	100.0	100.0	100.0	N/A	100.0
Mean	--	---	-0.03	-0.04	N/A	0.00
S.D.	--	---	0.05	0.04	N/A	0.06
Maximum	--	---	0.06	0.00	N/A	0.10
Minimum	--	---	-0.12	-0.11	N/A	-0.09

[a] Each item is identified in only one DIF category. If the item was flagged for more than one comparison analysis then the largest absolute DIF value indicates its category across all comparisons.

[b] N/A - Insufficient sample size (N < 200) for DIF analysis.

Table 11

Differential Item Functioning (DIF) Summary
For **MH D-DIF** on **SPR** Items of
Math Form B Using **Internal** or **External Criteria**
Spring Trials 1989 Administration

	Category of DIF Value For All Comparisons					
	CROSS-GROUP[a]	CROSS-GROUP[a]	MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISPANIC	WHITE/ ASIAN
MH D-DIF Category	Number	% of Items	Percent of Items by DIF Category			
		Internal Criterion				
DIF ≥ 1.5	1	0.0	0.0	0.0	10.0	0.0
1.0≤ DIF < 1.5	1	0.0	0.0	0.0	0.0	10.0
0.5≤ DIF < 1.0	0	0.0	0.0	0.0	0.0	0.0
0.0≤ DIF < 0.5	0	35.0	20.0	20.0	0.0	10.0
-0.5< DIF < 0.0	1	20.0	10.0	10.0	60.0	60.0
-1.0< DIF ≤-0.5	2	20.0	40.0	20.0	30.0	20.0
-1.5< DIF ≤-1.0	3	20.0	20.0	40.0	0.0	0.0
DIF ≤-1.5	2	25.0	10.0	10.0	0.0	0.0
Total	10	100.0	100.0	100.0	100.0	100.0
Mean	--	---	-0.69	-0.85	-0.23	-0.39
S.D.	--	---	0.57	0.64	0.69	0.57
Maximum	--	---	0.24	0.16	1.73	1.08
Minimum	--	---	-1.88	-2.00	-0.76	-1.02
		External Criterion				
DIF ≥ 1.5	0	0.0	0.0	0.0	N/A	N/A
1.0≤ DIF < 1.5	0	0.0	0.0	0.0	N/A	N/A
0.5≤ DIF < 1.0	0	0.0	0.0	0.0	N/A	N/A
0.0≤ DIF < 0.5	0	0.0	0.0	0.0	N/A	N/A
-0.5< DIF < 0.0	3	30.0	30.0	44.4	N/A	N/A
-1.0< DIF ≤-0.5	3	30.0	50.0	11.1	N/A	N/A
-1.5< DIF ≤-1.0	1	10.0	10.0	22.2	N/A	N/A
DIF ≤-1.5	3	30.0	10.0	22.2	N/A	N/A
Total	10	100.0	100.0	100.0	N/A	N/A
Mean	--	---	-0.66	-0.88	N/A	N/A
S.D.	--	---	0.43	0.71	N/A	N/A
Maximum	--	---	-0.22	-0.01	N/A	N/A
Minimum	--	---	-1.59	-2.15	N/A	N/A

[a] Each item is identified in only one DIF category. If the item was flagged for more than one comparison analysis then the largest absolute DIF value indicates its category across all comparisons.

[b] N/A - Insufficient sample size (N < 200) for DIF analysis.

Table 12

Differential Item Functioning (DIF) Summary
For STD P-DIF on SPR Items of
Math Form B Using **Internal or External Criteria**
Spring Trials 1989 Administration

			Category of DIF Value For All Comparisons			
	CROSS- GROUP[a]	CROSS- GROUP[a]	MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISPANIC	WHITE/ ASIAN
STD P-DIF Category	Number	% of Items	Percent of Items by DIF Category			
		Internal Criterion				
DIF ≥ .15	0	0.0	0.0	0.0	0.0	0.0
.10≤ DIF < .15	0	10.0	0.0	0.0	0.0	0.0
.05≤ DIF < .10	1	10.0	0.0	0.0	0.0	10.0
.00≤ DIF < .05	0	0.0	20.0	20.0	10.0	0.0
-.05< DIF < .00	4	40.0	50.0	40.0	80.0	60.0
-.10< DIF ≤-.05	3	30.0	20.0	30.0	10.0	30.0
-.15< DIF ≤-.10	2	20.0	10.0	10.0	0.0	0.0
DIF ≤-.15	2	0.0	0.0	0.0	0.0	0.0
Total	10	100.0	100.0	100.0	100.0	100.0
Mean	--	---	-0.05	-0.03	-0.02	-0.03
S.D.	--	---	0.04	0.03	0.02	0.04
Maximum	--	---	0.00	0.01	0.01	0.07
Minimum	--	---	-0.12	-0.11	-0.05	-0.07
		External Criterion				
DIF ≥ .15	0	0.0	0.0	0.0	N/A	N/A
.10≤ DIF < .15	0	0.0	0.0	0.0	N/A	N/A
.05≤ DIF < .10	0	0.0	0.0	0.0	N/A	N/A
.00≤ DIF < .05	0	0.0	0.0	10.0	N/A	N/A
-.05< DIF < .00	6	60.0	60.0	60.0	N/A	N/A
-.10< DIF ≤-.05	2	20.0	30.0	20.0	N/A	N/A
-.15< DIF <-.10	2	20.0	10.0	10.0	N/A	N/A
DIF <-.15	0	0.0	0.0	0.0	N/A	N/A
Total	10	100.0	100.0	100.0	N/A	N/A
Mean	--	---	-0.04	-0.04	N/A	N/A
S.D.	--	---	0.03	0.04	N/A	N/A
Maximum	--	---	0.00	0.00	N/A	N/A
Minimum	--	---	-0.11	-0.12	N/A	N/A

[a] Each item is identified in only one DIF category. If the item was flagged for more than one comparison analysis then the largest absolute DIF value indicates its category across all comparisons.

[b] N/A - Insufficient sample size (N < 200) for DIF analysis.

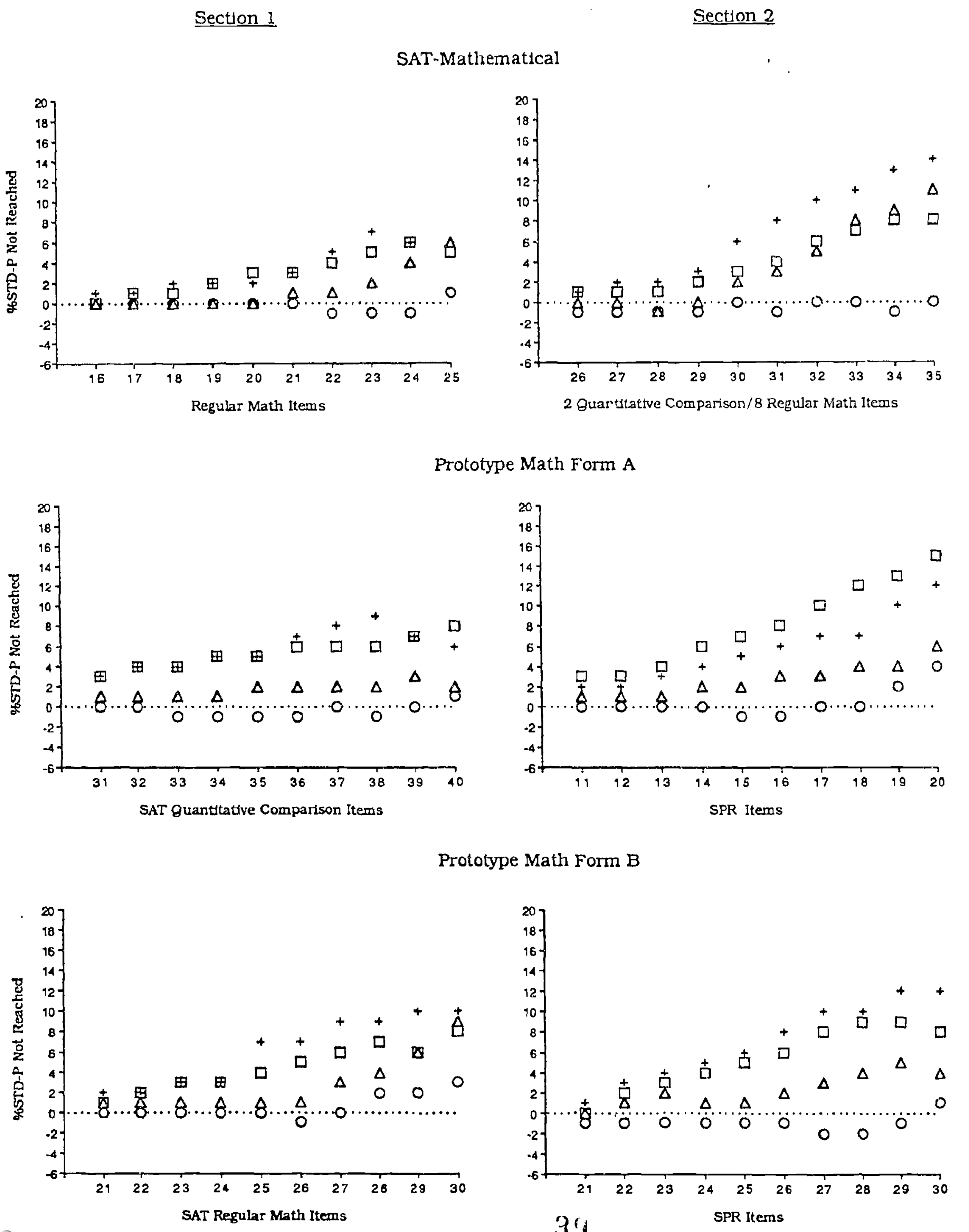
Table 13

Differential Speededness Summary
for SAT-M and Math Prototypes A and B
Last 10 Items of Each Section
Spring Trials 1989 Administration

STD P-DIF Not Reached Category	Category of Maximum Absolute DIF Value For All Comparisons[1]		Comparison			
			MALE/ FEMALE	WHITE/ BLACK	WHITE/ HISP.	WHITE/ ASIAN
	Number	% of Items	Percent of Items by DIF Category			
SAT-M, Section 1 (10 Regular Math items)						
DIF ≤ -.10	0	0.0	0.0	0.0	0.0	0.0
-.10 < DIF ≤ -.05	0	0.0	0.0	0.0	0.0	0.0
-.05 < DIF < 0.05	6	60.0	100.0	70.0	60.0	90.0
0.05 ≤ DIF < 0.10	4	40.0	0.0	30.0	40.0	10.0
DIF ≥ 0.10	0	0.0	0.0	0.0	0.0	0.0
SAT-M, Section 2 (2 Quantitative Comparison, 8 Regular Math items)						
DIF ≤ -.10	0	0.0	0.0	0.0	0.0	0.0
-.10 < DIF ≤ -.05	0	0.0	0.0	0.0	0.0	0.0
-.05 < DIF < 0.05	4	40.0	100.0	60.0	40.0	60.0
0.05 ≤ DIF < 0.10	2	20.0	0.0	40.0	20.0	30.0
DIF ≥ 0.10	4	40.0	0.0	0.0	40.0	10.0
Prototype Math Form A, Section 1 (10 Quantitative Comparison items)						
DIF ≤ -.10	0	0.0	0.0	0.0	0.0	0.0
-.10 < DIF ≤ -.05	0	0.0	0.0	0.0	0.0	0.0
-.05 < DIF < 0.05	3	30.0	100.0	30.0	30.0	100.0
0.05 ≤ DIF < 0.10	7	70.0	0.0	70.0	70.0	0.0
DIF ≥ 0.10	0	0.0	0.0	0.0	0.0	0.0
Prototype Math Form A, Section 2 (10 SPR items)						
DIF ≤ -.10	0	0.0	0.0	0.0	0.0	0.0
-.10 < DIF ≤ -.05	0	0.0	0.0	0.0	0.0	0.0
-.05 < DIF < 0.05	3	30.0	100.0	30.0	40.0	90.0
0.05 ≤ DIF < 0.10	3	30.0	0.0	30.0	40.0	10.0
DIF ≥ 0.10	4	40.0	0.0	40.0	20.0	0.0
Prototype Math Form B, Section 1 (10 Regular Math items)						
DIF ≤ -.10	0	0.0	0.0	0.0	0.0	0.0
-.10 < DIF ≤ -.05	0	0.0	0.0	0.0	0.0	0.0
-.05 < DIF < 0.05	4	40.0	100.0	50.0	40.0	80.0
0.05 ≤ DIF < 0.10	4	40.0	0.0	50.0	40.0	20.0
DIF ≥ 0.10	2	20.0	0.0	0.0	20.0	0.0
Prototype Math Form B, Section 2 (10 SPR items)						
DIF ≤ -.10	0	0.0	0.0	0.0	0.0	0.0
-.10 < DIF ≤ -.05	0	0.0	0.0	0.0	0.0	0.0
-.05 < DIF < 0.05	3	30.0	100.0	40.0	30.0	90.0
0.05 ≤ DIF < 0.10	3	30.0	0.0	60.0	30.0	10.0
DIF ≥ 0.10	4	40.0	0.0	0.0	40.0	0.0

[1] Each item is identified in only one DIF category. If the item was flagged for more than one comparison analysis then the largest absolute DIF value indicates its category across all comparisons.

Figure 1: Mathematical Tests Differential Speededness
Spring Trials 1989 Administration
Last 10 Items of Each Section


○ Male/Female
□ White/Black
+ White/Hispanic
△ White/Asian American
Section 1
Section 2
SAT-Mathematical
%STD-P Not Reached
Regular Math Items
2 Quantitative Comparison/8 Regular Math Items
Prototype Math Form A
SAT Quantitative Comparison Items
SPR Items
Prototype Math Form B
SAT Regular Math Items
SPR Items