DOCUMENT RESUME

ED 247 260 TM 840 431

AUTHOR Kulick, Edward; Dorans, Neil J.
TITLE The Standardization Approach to Assessing Unexpected Differential Item Performance.
PUB DATE Apr 84
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Entrance Examinations; Individual Differences; *Mathematical Models; Performance Factors; Sample Size; *Statistical Analysis; *Test Bias; *Test Items
IDENTIFIERS Outliers; Scholastic Aptitude Test

ABSTRACT

A new approach to assessing unexpected differential item performance (item bias or item fairness) is introduced and applied to the item responses of different subpopulations of Scholastic Aptitude Test (SAT) takers. The essential features of the standardization approach are described. The primary goal of the standardization approach is to control for differences in subpopulation ability before making comparisons between subpopulation performance on test items. By so doing, it removes the contaminating effects of ability differences from the assessment of item fairness. The approach is capable of identifying rare individual instances (outliers) of unexpected differential item performance (that can sometimes be attributed to unfair content), as well as differences on groups of items which might be attributed to the fact that these items are measuring different attributes in different subpopulations. (Author)

The Standardization Approach to Assessing Unexpected Differential Item Performance[1]

Edward Kulick
Neil J. Dorans

Educational Testing Service

[1] A paper presented at the annual meeting of AERA in New Orleans, April 1984.

The Standardization Approach to Assessing Unexpected Differential Item Performance

Edward Kulick
Neil J. Dorans

Precis

A new approach to assessing unexpected differential item performance (item bias or item fairness) is introduced and applied to the item responses of different subpopulations of Scholastic Aptitude Test (SAT) takers. The essential features of the standardization approach are described. The primary goal of the standardization approach is to control for differences in subpopulation ability before making comparisons between subpopulation performance on test items. By so doing, it removes the contaminating effects of ability differences from the assessment of item fairness. The approach is shown to be capable of identifying rare individual instances (outliers) of unexpected differential item performance (that can sometimes be attributed to unfair content), as well as differences on groups of items which might be attributed to the fact that these items are measuring different attributes in different subpopulations.

The Standardization Approach to Assessing
Unexpected Differential Item Performance

Introduction

In recent years much attention has been directed to the issue of fairness in educational and psychological tests. At Educational Testing Service (ETS), those who develop and review the Scholastic Aptitude Test/Test of Standard Written English (SAT/TSWE) are aware of the diversity of the test-taking population and attempt to construct tests based on a broad sampling of tasks and topics that tend not to favor any subgroup of the population. In addition, there are a number of procedures which ETS has instituted (Donlon, 1981), including sensitivity reviews and statistical checks, in order to guard against possible favoritism on the SAT towards any subgroup. Nevertheless, despite these efforts, the importance and complexities inherent in the nature of item fairness necessitate post hoc investigations to evalute the effectiveness of these safeguarding procedures. This paper summarizes the findings of four studies that used the statistical method of standardization to examine whether there are unexpected differences in item performance across different subpopulations of the SAT test-taking population. In addition, a brief introduction to the standardization method is presented.

Standardization Methodology

An item is exhibiting unexpected differential item performance when the probability of correctly answering the item is lower for examinees from one group than for examinees of equal ability from another group or groups. This definition may be formalized mathematically by letting S represent ability as measured by total score on the standard College Board 200-to-800 SAT scale (or on the 20-to-60 TSWE scale), and X represent an item score (1 if the answer to the question is correct and 0 if the answer is incorrect). An item,

then, is free of unexpected differential item performance when it satisfies the following equality

$$P_g(X=1|S)=P_{g'}(X=1|S) \quad \text{for all subpopulations g and g',}$$

where $P_g(X=1|S)$ is defined as the probability that candidates from subpopulation g who have total test scores equal to S will answer the item correctly. For example, if male and female candidates with the same total test scores do not have equal probabilities of successful performance on the item, this difference in probabilities is taken as evidence of unexpected differential item performance for male and female candidates at this score level. Note that a lack of unexpected differential item performance does not imply that there will not be any observed differences in item performance across subgroups of the SAT candidate population, but that there are no differences in conditional item performance across subgroups when the requisite condition before comparison is identical total test score. The reference to this type of differential performance as "unexpected" is purposeful, in order to emphasize that the focus ought to be on differences between candidates of equal score level, among whom one would not expect to find any differences. This represents an important distinction from observed differences in item performance between groups of varying ability, where some differences are of course expected.

Previous methods used to appraise unexpected differential item performance typically have been hampered by sensitivities to differences in overall subpopulation ability or differences in item quality (discrimination). The standardization methodology, however, controls for differences in both subpopulation ability and in item quality. Standardization is used here to

mean that differences on one variable have been controlled for prior to making comparisons between groups on some other related variable. A general approach to assessing unexpected differences in item performance via standardization is described in detail in Dorans and Kulick (1983). The essential features of the method as applied to the SAT are as follows: Using the standard College Board 200-800 SAT scale one can establish 61 individual ability levels (200, 210, 220, etc.). The probability that an examinee at a given ability level will correctly answer an item can be estimated by the observed percent correct among those with the given scaled score. Studies of unexpected differential item performance focus on differences between two or more groups. One group is arbitrarily designated as the base group. The base group is used to estimate the conditional probability of successful item performance given score level. Usually the group that provides the most stable estimates of the conditional probabilities across the entire scaled score range is selected as the base group. Typically, but not always, this is the largest group. The remaining groups are referred to as study groups, or comparison groups.

Several indices used in the standardization process may be defined. P_b is the overall percent correct in the base group for an item. P_{bs} is the percent correct at ability level s in the base group. P_g is the overall percent correct in the study group. P_{gs} is the percent correct at ability level s in the study group. P_b and P_g are not directly comparable when the base group and study group have different marginal ability distributions. It is necessary to calculate the expected item performance of the study group, $\hat{P}_g$. $\hat{P}_g$ is computed by taking a weighted sum of the 61 conditional probabilities of successful item performance observed in the base group, P_{bs}, where the

relative frequencies at each of the 61 scaled score levels in a designated group serve as the weights. The designated group that supplies the frequency distribution to be used as weights is referred to as the standardization group. Having the study group also serve as the standardization group (as was done in the four studies presented here) insures that the most important conditional probabilities are weighted most heavily, i.e., conditional probabilities at those score levels most attained by the study group.

The most precise measure of differential item performance is at the individual scaled score level, $D_{gs} = P_{gs} - P_{bs}$. These differences can be combined across score levels in a variety of ways to obtain a number of summary indices of unexpected differential item performance. Plots of these differences, as well as plots of P_{gs} and P_{bs} are helpful to visualize the quantification of unexpected differential item performance (see Figures 1-4). Figures 1 and 2 depict an item that is performing fairly for both groups. Figures 3 and 4 portray an item that is unexpectedly difficult for females. The top figures (1 and 3) present the conditional probabilities of successful item performance for males and females. These curves may also be thought of as nonparametric item-test regressions or empirical item characteristic curves. The lower Figures (2 and 4) are simply plots of the group differences observed above.

One of the most informative indices summarizing these differences is the root mean weighted squared difference ($RMWSD_g$). The $RMWSD_g$ for an item is obtained by squaring each difference in conditional probabilities of successful item performance between the study and base groups, D_{gs}, taking a weighted sum of these squared differences, and taking the square root of the

weighted sum, where the relative frequency distribution of the standardization group serves as the weighting function. Since this index is unsigned, any difference produces a positive discrepancy. Consequently, every item will have a non-negative value of $RMWSD_g$. An item exhibiting substantial unexpected different item performance will have a large $RMWSD_g$. An item exhibiting absolutely no unexpected differential item performance will have a $RMWSD_g$ equal to zero.

The difference (D_g) between P_g and $\hat{P}_g$, ($D_g = P_g - \hat{P}_g$), is another index of unexpected differential item performance. If there is no unexpected differential item performance between the study group and base group, D_g should equal zero. A positive D_g indicates that the study group exceeds its expected performance, while a negative D_g indicates that the item is harder than expected for the study group.

A problem faced by any investigation which seeks to detect and quantify unexpected differential item performance, regardless of methodology, is the determination of what level of unexpected differential item performance should evoke concern. In the first report using the standardization approach (Dorans and Kulick, 1983), an empirical determination was made concerning the practical cutoff point for values of $RMWSD_g$ using frequency distributions of the $RMWSD_g$ index. According to this determination, an item with a $RMWSD_g$ greater than or equal to .08 merits careful investigation, while an item with a $RMWSD_g$ less than .08 does not require additional study. Items with $RMWSD_g$ greater than or equal to .16 are exhibiting clearly unacceptable levels of differential performance. Figure 5 presents a plot of the $RMWSD_g$ index for a set of verbal items. The value of $RMWSD_g$ equals the distance from the origin

to the point representing the item. Projection of each point on the horizontal axis yields D_g for that item. Most of the items in this figure fall within the smallest arc. One item, however, can be seen falling outside the second arc. This is clearly an outlier exhibiting a high level of unexpected differential item performance.

Results Using the Standardization Method

Four studies have been completed to date employing the standardization approach to item bias. The findings from these studies are briefly summarized below.

The first investigation compared the performance of male and female candidates on a form of the SAT administered in 1977. Essentially there was very little evidence of unexpected differential item performance. Figure 5 shows the distribution of $RMWSD_f$ values on the verbal test. A few items are in the region where they should be examined more closely, but the most striking feature of the plot is the analogy outlier. Clearly this item is exhibiting an unacceptable level of unexpected differential item performance. This same item is portrayed in Figures 3 and 4. Notice the largest differences are at the lower to middle portion of the scaled score range, where the majority of the candidates are. Examination of this item revealed that a certain knowledge of hunting and fishing are required to answer correctly. It should be noted that this form of the SAT was developed prior to the institution of formal sensitivity reviews.

The second study divided the candidate population into three subgroups based on reported level of fathers' education, a variable related to

socioeconomic status. The education levels defining the first study group, second study group, and base group were: less than high school degree, high school degree but less than bachelor's degree, and bachelor's degree or higher, respectively. Thus each item was evaluated twice, once with respect to each study group, while maintaining the same base group.

Examination of discrepancy index summary statistics revealed that there was little evidence of systematic unexpected differential item performance by either study group on SAT-V, SAT-M or TSWE. The same conclusion was reached by inspection of frequency distributions and plots of item discrepancy indices such as the one in Figure 6. The results of this study seem to indicate that the items on the SAT and TSWE forms used in this study are equally appropriate for all candidates regardless of father's level of education.

The third study divided the candidate population into two subgroups based on reported answers to a racial/ethnic background question. The Oriental group (including Asian Americans and Pacific Islanders as well) was designated as the study group, while the White (or Caucasian) group served as the base group. Whereas studies I and II had found few or no outliers, this investigation detected 52 (out of 195) items which displayed questionable levels of unexpected differential item performance. Figure 7 indicates clearly that unexpected differential item performance between Oriental and White candidates was rather widespread on this particular mathematical test form. Similar plots were observed for SAT-V and TSWE.

Two factors were identified which may help account for the abundance of items identified: 1) since a sizeable percentage of the Oriental group reported that English is not their best language, it was suggested that items

covering verbal skills which this subgroup had not mastered would appear unduly difficult for them; and 2) the sample size obtained for the Oriental group may have been too small to accurately estimate conditional percents correct. The language hypothesis was tested on the mathematics set of items. A test developer independently divided the math items into categories of "verbally-loaded" math items, "pure" math items, and "neutral" math items. Analysis of the discrepancy indices on items in each category supported the explanation we proposed, as the "verbally-loaded" category had the most unexpectedly difficult items for the Oriental group, while the "pure" math category had the most unexpectedly easy items for the Oriental group. The effects of small sample size combined with the heterogeneous composition of the Oriental sample on the non-parametric item-test regression curves is apparent in Figure 8. Observe the erratic pattern of stars in this plot.

This study demonstrates that in situations where the test becomes multi-dimensional for one of the groups, the scaled score may not be an effective control variable. These results suggest that further investigations of SAT/TSWE items need to be done where the Oriental group is restricted to those for whom English is the best language.

The fourth and final study divided the candidate population into two subgroups based on reported answers to a racial/ethnic background question. The Black group was designated as the study group, while the White group served as the base group. Examination of discrepancy index summary statistics at the item type level revealed an interesting finding: Analogy type items appeared to be unexpectedly more difficult for Blacks than for Whites. Since this result is consistent with previous research on the SAT (see Dorans (1982)

for a review), and is not readily explainable, it suggests the need for additional research in order to determine possible factors or characteristics of the analogy type items which may be related to ethnicity. Further analyses revealed that the test, as a whole, was relatively free from unexpected differential item performance between Blacks and Whites. Most evidence of unexpected differential item performance was limited to a few items, and only one of these exhibited a clearly unacceptable level. The nonparametric item-test regressions for this item (and their differences) are presented in Figures 9 and 10. Inspection of the item content provided no insight to account for the differential performance observed on the item. Additional analyses and examination of the item by test development staff are recommended.

In sum, the standardization method seems to be an effective means of comparing the item performance of groups who differ greatly in ability. Its major drawback is probably the large sample sizes that it requires, but for its current application to the SAT/TSWE this is not a serious weakness. Furthermore, the visual displays that it provides, both at the item and test level, are valuable aides to data interpretation.

References

Donlon, T. F. The SAT in a diverse society: Fairness and sensitivity. The College Board Review, No. 122 (Winter 1981-82), 16-21, 30-32.

Dorans, N. J. Technical review of item fairness studies: 1975-1979 (SR-82-90). Princeton, NJ: Educational Testing Service, 1982.

Dorans, N. J. and Kulick, E. Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach, ETS Research Report (RR-83-9). Princeton, NJ: Educational Testing Service, 1983.

Kulick, E. Assessing unexpected differential item performance of Black candidates on SAT form CSA6 and TSWE form E33, ETS Statistical Report (in press). Princeton, NJ: Educational Testing Service, 1984.

Kulick, E. and Dorans, N. J. Assessing unexpected differential item performance of candidates reporting different levels of father's education on SAT form CSA2 and TSWE form E29, ETS Statistical Report (SR-83-27). Princeton, NJ: Educational Testing Service.

Kulick, E. and Dorans, N. J. Assessing unexpected differential item performance of Oriental candidates on SAT form CSA6 and TSWE form E33, ETS Statistical Report (SR-83-106). Princeton, NJ: Educational Testing Service.

Conditional Probabilities of Successful Item Performance
for Males and Females on Two Verbal Items from SAT Form ZSA5

Figure 1

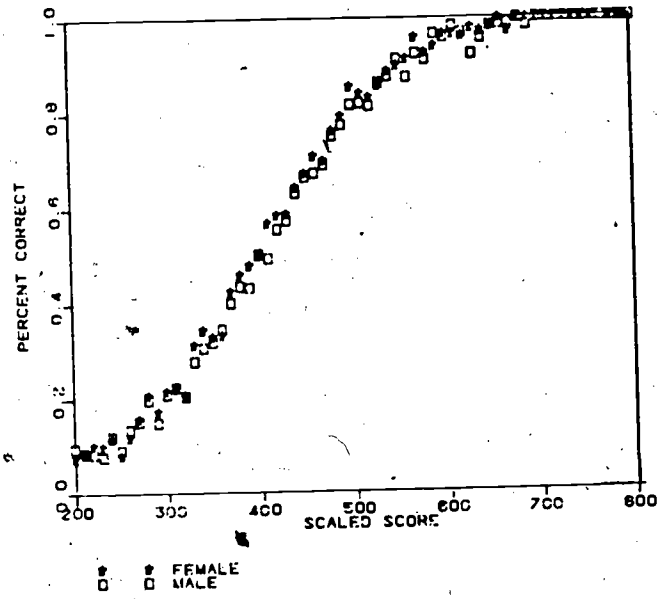
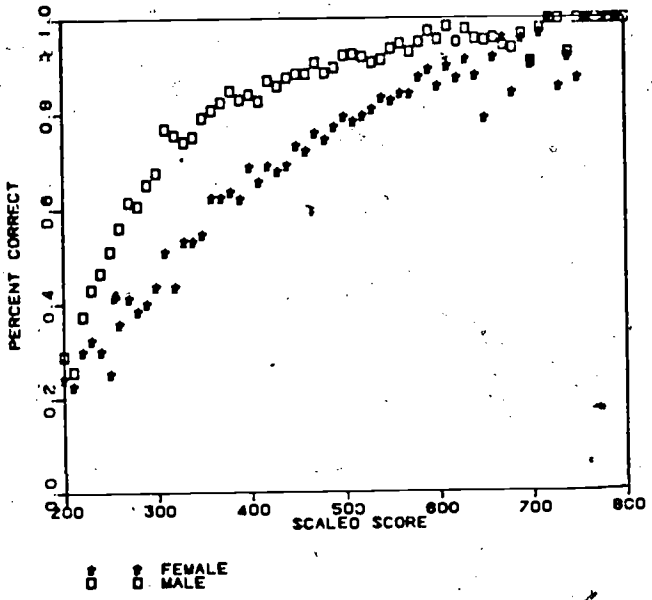


Figure 3



Difference Plots of Two Verbal Items from SAT Form ZSA5

Figure 2

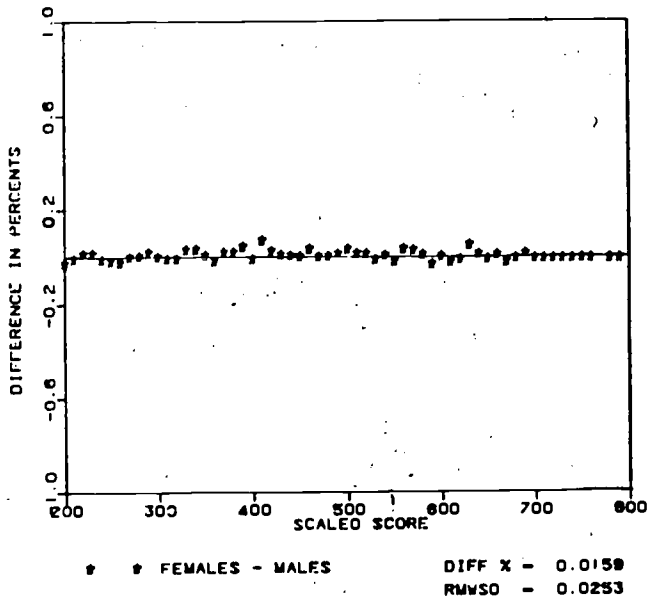


Figure 4

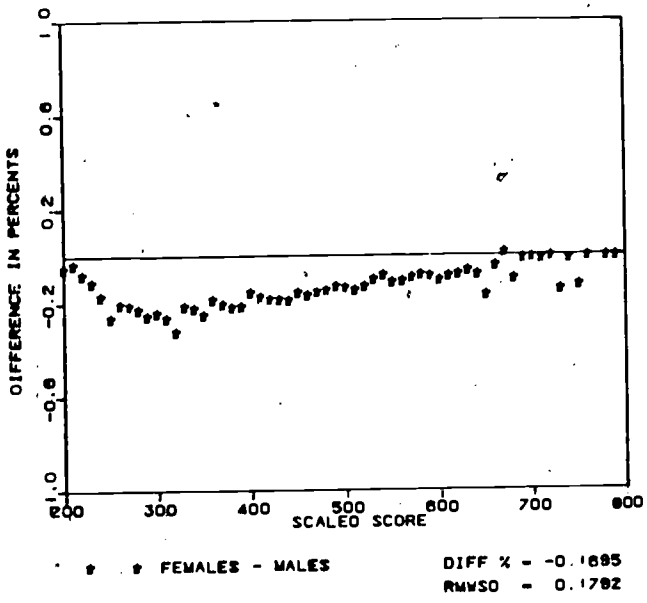
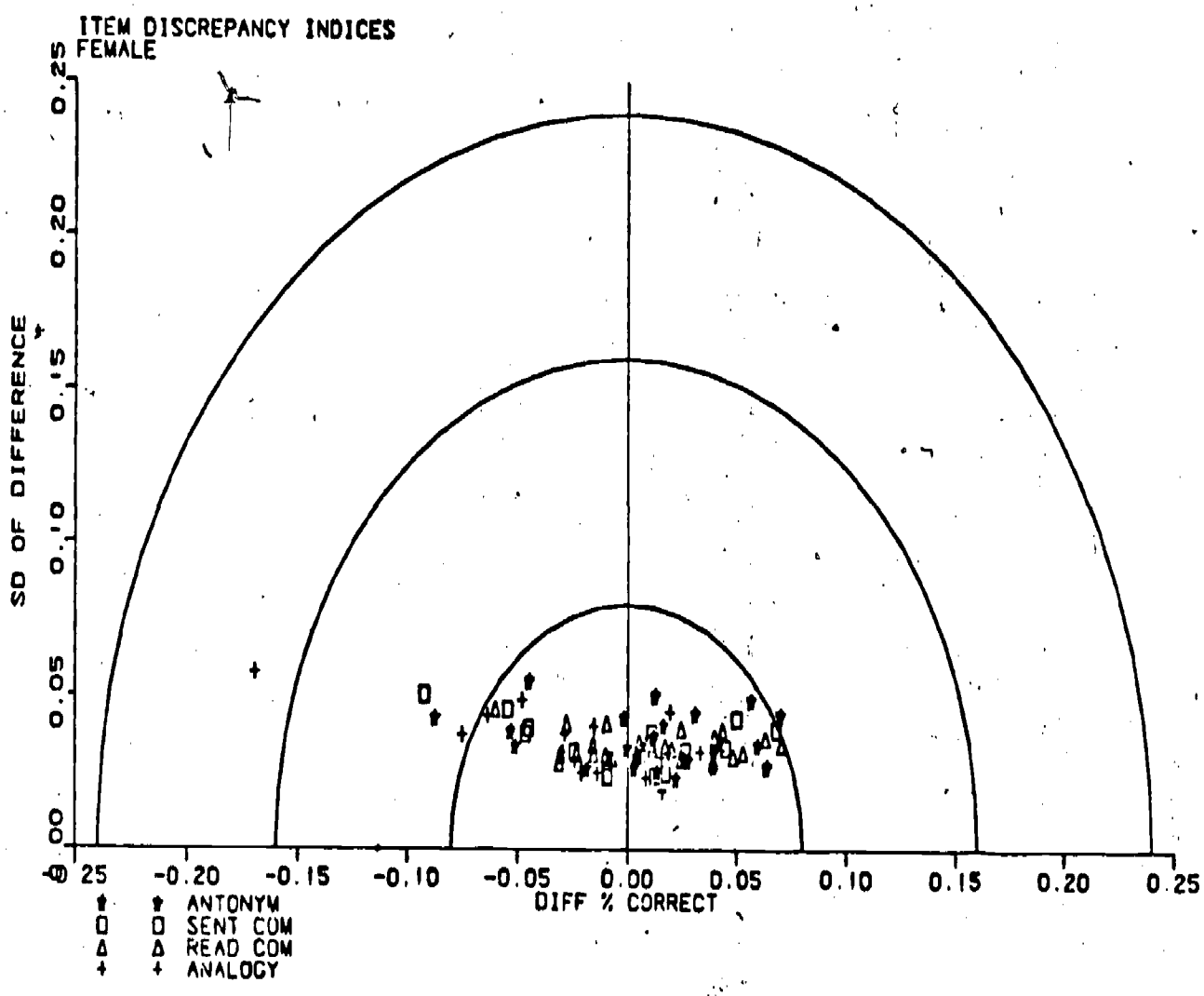


Figure 5

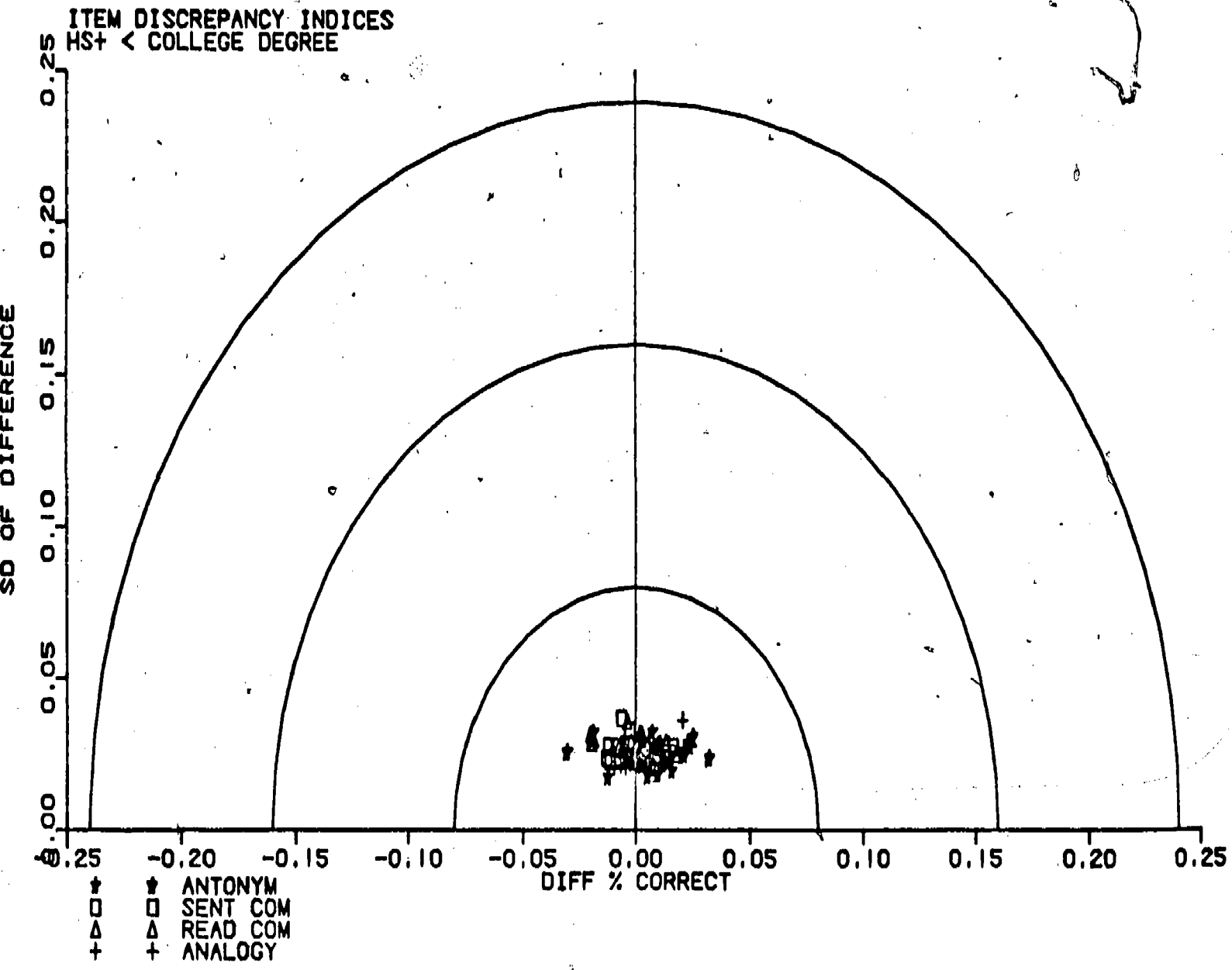
Plot of Root Mean Weighted Squared Differences (RMWSD)[a] Between the Conditional Probabilities of Success for Male and Female Candidates on Verbal Items from SAT Form ZSA5



[a]RMWSD equals the distance from the origin to the point representing the item. Projection of each point on the horizontal axis yields the difference between P_y and $\hat{P}_y$, D_y, for that item. Projection of each point on the vertical axis yields the standard deviation of the weighted differences, an idex of residual crossover.

Figure 6

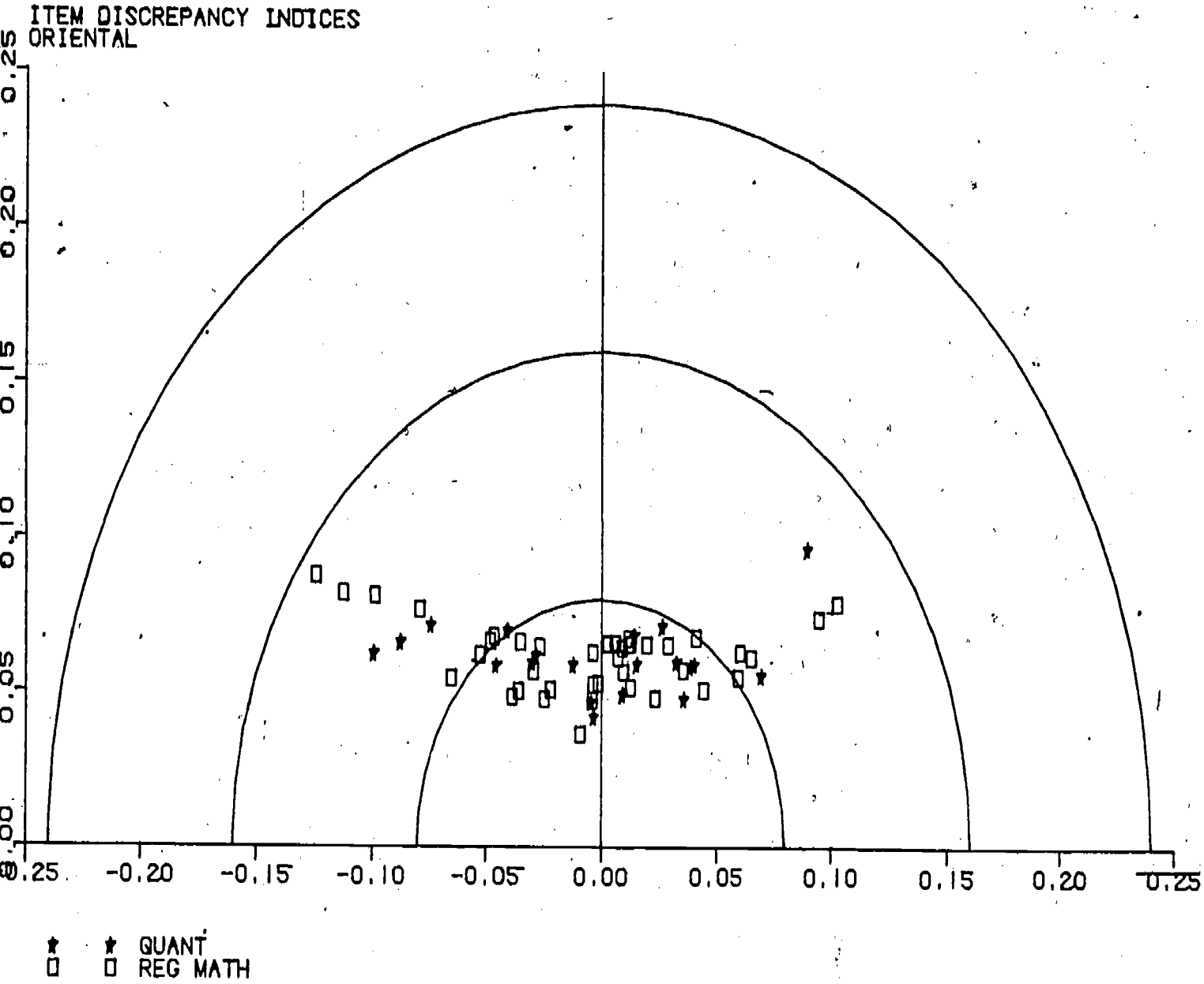
Plot of Root Mean Weighted Squared Differences (RMWSD[a]) Between the Conditional Probabilities of Success for Study Group 2 and the Base Group on Verbal Items from SAT Form CSA2



[a]RMWSD equals the distance from the origin to the point representing the item. Projection of each point on the horizontal axis yields the difference between P_y and $\hat{P}_y$, D_y, for that item. Projection of each point on the vertical axis yields the standard deviation of the weighted differences, an idex of residual crossover.

Figure 7

Plot of Root Mean Weighted Squared Differences (RMWSD[a]) Between the Conditional Probabilities of Success for Orientals and Whites on Verbal Items from SAT Form CSA6



[a]RMWSD equals the distance from the origin to the point representing the item. Projection of each point on the horizontal axis yields the difference between P_y and $\hat{P}_y$, D_y, for that item. Projection of each point on the vertical axis yields the standard deviation of the weighted differences, an index of residual crossover.

Figure 8

Example of Variability in the Conditional Probabilities of Successful Item Performance for Orientals on a Math Item from SAT Form CSA6

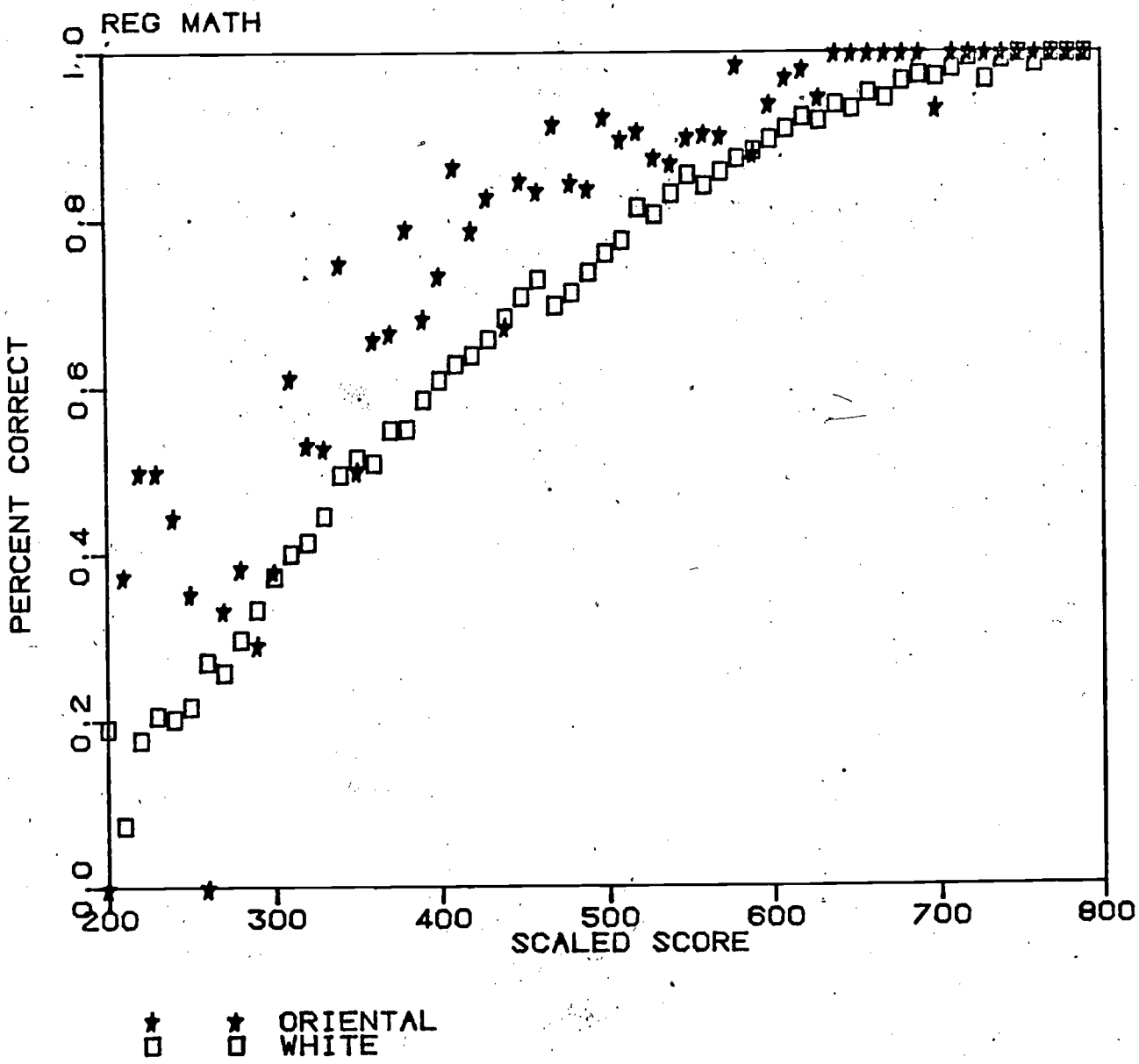


Figure 9

Conditional Probabilities of Successful Item Performance for Blacks and Whites on an Analogy Item from SAT Form CSA6

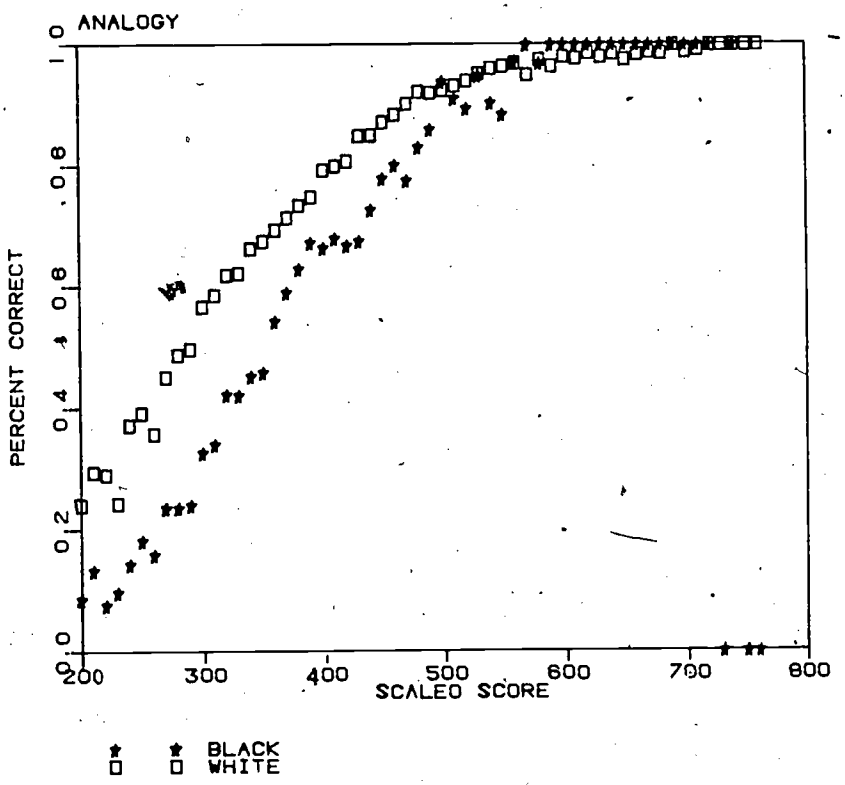


Figure 10

Differences Between Conditional Probabilities