DETECTING POTENTIALLY BIASED TEST ITEMS

John R. Hills
&
F.J. King
Florida State University

INTRODUCTION

In the development of the Florida Statewide Assessment Tests, items are reviewed by several different panels of judges with the intention of removing or modifying any items which might be biased against or in favor of major recognizable cultural groups. Even with that careful attention to this problem during test development, the mean scores for different cultural groups can be widely different. Examination of the items for bias after pretest or after the official administration is often suggested as a means of detecting and removing the influence of any biased items which might have passed through the careful screening.

The purpose of this study was to test several possible simplifications of the transformed item difficulty (TID) technique for assessing potential for item bias of items in the Florida Statewide Assessment Tests. In a previous unpublished paper, Hills and King (1982) recommended the use of the TID method but with very large samples and a sampling procedure for matching for ability that has now been superseded. They also matched for ability on the measure being studied. In this study much smaller samples were used, along with the new sampling procedure available in SPSS. Results were compared, matching for ability on both the measure being studied and the other available measure, communications or mathematics score on SSAT II.

DEFINITION OF ITEM BIAS

An important distinction must be made between the terms "potentially biased" and "biased." The difference is that an item may be easier for one group than another without being biased. One group may have achieved more than the other. So to locate bias in item difficulty, one must first arrange so that the two groups being compared have equal achievement levels (Hunter, 1975).

Now, if groups of equal achievement levels have distinctly different item *p* values, that may be evidence of bias in the item. Bias, however, is defined as a difference in proportion correct due to irrelevant influences. If the difference in proportion correct reveals a real difference in relevant achievement between the two groups, the item is doing exactly what it should be doing and cannot reasonably be called biased.

The use of statistical procedures to isolate items that yield different probabilities of correct response for students of different groups when the groups have been equated for achievement level, is, therefore, merely a search for potentially-biased items. The items which display differences in probabilities of correct responses may be revealing real differences in achievement, or the differences may be unexplainable. But the only items properly regarded as biased are those in which it is clear that an identifiable, irrelevant aspect caused the difference between the groups' performance. An identifiable, irrelevant aspect means that people who are sensitive to the content of an item and to the relevant cultures of the groups being compared will agree in selecting the item as one which will yield a difference in success rates. Further, they must agree on the reason for the difference. Finally, the reason they give for the difference must be irrelevant to the subject matter being measured.

THE TRANSFORMED ITEM DIFFICULTY (TID) METHOD

There are many ways to attempt, statistically, to detect potentially-biased items. At least 50 methods or variations on methods are currently discussed in the literature. An old standby is the TID method. TID stands for transformed item difficulty; users often transform item _p_ values into normal curve deviates, and then transform those normal curve deviates into a particular scale used by the College Entrance Examination Board and Educational Testing Service for their internal analyses. The TID label has come to stand for any method that concentrates on overall item _p_ value comparisons. Some other methods examine performance of cultural groups at different ability levels, some examine the item characteristic curves for an item used with different cultural groups, and so on. No method is generally superior to the others, and, unfortunately, the methods do not all agree on either the number of potentially-biased items or on which items are potentially biased.

The simple comparisons of item _p_ values, untransformed, is a useful approach for the Statewide Assessment Tests if the cultural groups are matched for ability. The tests are quite easy, being minimum competency tests, and differences between groups, both of which have above 90% passing rates on an item, are not going to be of practical significance. The normal curve transformation only serves to emphasize differences at the extremes of the _p_ value scale, and there are essentially no items at the low end of that scale in these tests. We will call the method the ID approach since it uses item difficulties which are untransformed.

Two assumptions made by the TID and ID methods are that the test is unidimensional, i.e., all items measure the same dimension, and that the groups being compared for evidence of potential item bias are of equal achievement levels. The assumption of unidimensionality is a problem because experts do not agree on methodology for testing the assumption. For our current purposes, the recommendations of Reckase (1979) for determining unidimensionality of a test were followed: (1) In a factor analysis of test items, the first unrotated principal component accounts for at least 20% of the test variance; (2) The first principal component is large relative to the other factors. When the two groups differ in achievement level, one way to meet the second assumption of equal achievement is to select randomly from the larger group a subset equal in achievement to the smaller group.

Prior work by the authors tested the unidimensionality assumption by extracting principal components from item correlation matrices computed on randomly selected samples of 1000 eleventh grade students. Twenty six communication items and eight mathematics items were eliminated because they were answered correctly by at least 95% of the subjects. The proportions of variance accounted for by the first principal component, for both communication and mathematics, equalled or exceeded 20%, and the first principal components were at least 4 times as large as the next largest factors. Thus the criteria for unidimensionality as recommended by Reckase (1979) were judged to have been met.

The Hills and King (1982) study started with randomly selected samples of 20,000 whites, 7,300 Hispanics, 2000 blacks, 1000 males and 1000 females. Achievement was equalized to a satisfactory degree by the procedure described earlier of selecting, at random, within score intervals, subjects from the larger group equal to the number in the smaller group. The test score distributions were negatively skewed; 11 score intervals were used. The lowest two intervals each contained approximately 5% of the smaller group; each of the others contained about 10%. The intervals set for the smaller group were then used for both groups in ethnic comparisons, selection of cases taking place from the larger group. For example, Table 1 contains the score distributions for blacks and whites on SSAT II, Communications. In the first score interval, from a score of one to a score of 36, there are 101 blacks and 129 whites. This is about 5% of the blacks. Similarly, in the second score interval, 37 to 43, there are 100 blacks and 111 whites, about 5% of the blacks. In the third interval, 44 to 49, there are 177 blacks and 229 whites, about 10% of the blacks. At the other end, interval eleven includes only the score of 60, with 135 blacks and 5635 whites. The number of blacks is as close to 10% as any interval could yield. To equalize the groups in ability, 28 whites, chosen at random from those with scores from 1 to 36 were removed. As a result the first interval includes the same number of whites as blacks. Each interval is handled the same way. In the eleventh interval, 5500 randomly-selected whites were removed.

In the case of the male-female comparisons, neither group was consistently larger. Therefore, a comparison of group size was made at each score interval; selection was made from the group with the most cases in that score interval. After selection, the interval contained the same number of cases in each group. Approximately 2000 cases were used in each group for black-white comparisons, 1000 in each for male-female comparisons, and about 7300 in each for Hispanic-white comparisons.

After the groups were equated for achievement, proportion correct ($\underline{p}$) values were computed for each group for each item in each test. In general, the $\underline{p}$ values were very high for the communications test items, many of them being above .90 and a few being below .50. On the mathematics tests, however, there was wider range of difficulty, with fewer above .90 and some well below .50.

Table 1
Frequency Distributions
for Approximately 20,000 Whites and 2000 Blacks
on SSAT, Communications

Interval			
Number	Lower Limit	Black	White
1	1	101	129
2	37	100	111
3	44	177	229
4	50	186	407
5	53	240	578
6	55	162	615
7	56	171	1052
8	57	201	2014
9	58	235	3425
10	59	243	5499
11	60	135	5635

For each pair of groups being compared (whites vs. blacks, whites vs. Hispanics, and males vs. females), the item _p_ values for each score (communications and mathematics on SSAT I and SSAT II) were plotted by computer. (See Appendix A for examples of two such plots.) On each of the twelve plots, a line was drawn visually through the center of the data points from lower left to upper right to identify outlier items which might be biased. Parallel lines were then drawn equidistant on either side of the central axis. The distance was chosen so that no more than ten or so items in any plot would be selected as discrepant (outliers) by virtue of being outside those parallel lines. Where possible, such lines were drawn through an area of the plot which included few or no items. It was noted at this time that there was no clear evidence of curvilinearity of these plots.

PROCEDURES OF THIS STUDY

The specific purpose of the present study was to determine whether statistical results similar to those previously obtained by the black-white comparisons would occur when (1) the sample sizes were reduced, (2) the number of ability intervals decreased, and (3) ability matching was done on a variable other than the one whose items were being studied.

The procedure involved sampling, using the new SPSS Sample command, from the previous sample of 20,000 whites and 2000 blacks. A random sample of 5000 whites and 500 blacks was first obtained. Samples from those samples were then drawn in order to obtain random samples of 1000 whites and 100 blacks. (It ordinarily requires at least 10 times as many whites as blacks to find enough matches at the low end of the scale.) The white sample was matched to the black sample by forming intervals of approximately 20% of the black cases, noting the score values at the interval limits, and then sampling equal numbers from the white cases in those score intervals. The p values were then obtained for each group on the matched samples. Eight plots, four for each group size comparison, were obtained, one each for communications and math item when achievement matching was done on communications total scores and one for each set of items with achievement matching on mathematics. The items that were discrepant in plots of p values for these samples were compared with the discrepant items in the earlier plots which used samples of 2000 cases.

RESULTS

The results appear in Table 2. In that table, each row provides results for a comparison of outliers on the plot of 2000 whites vs. 2000 blacks matched on the skill being measured with the outliers on the smaller group matched as described in the figure. The top four rows are for smaller groups of 500 whites vs. 500 blacks. The bottom four rows are for groups of 100 whites vs. 100 blacks. In each set of four rows, the upper two rows are for data in which matching is on the variable being measured; the bottom two are for data in which matching is on

Table 2

Comparison of Outliers in Large Group (N=2000)

with Smaller Groups (N=500 and N=100)

		Outliers in:		
		Large and Small	Large not Small	Small not Large
N=500	Communication, Matched on Communications	4	3	1
	Mathematics Matched on Mathematics	8	2	4
	Communication Matched on Mathematics	4	4	3
	Mathematics Matched on Communication	7	4	5
N=100	Communication Matched on Communication	3	4	2
	Mathematics Matched on Mathematics	4	3	5
	Communication Matched on Mathematics	4	3	6
	Mathematics Matched on Communication	1	6	7

mathematics scores when communication is being measured or on communications scores when mathematics is being measured. The columns provide for entries showing how many items were outliers on the plots for both the large groups (N's of 2000 for each race) and the small groups, how many were outliers in the plot for the large groups but not the small groups, and how many were outliers for the small groups but not the large groups.

It can be seen in the plot that there is roughly a descending order of agreement between plots. When the groups are larger--500 instead of 100--more outliers are common to both plots and fewer are outliers in one but not the other. In the plots of 500, 23 items are outliers in both group sizes, and 26 are outliers in one plot but not the other. In the groups of 100, only 12 items are outliers in plots of large and small groups, and 39 are outliers in one group but not the other. Thus, one can conclude that small groups are not satisfactory for deciding which items are potentially biased, but when group sizes reach 500 of each group or so, there is a reasonable amount of agreement with the outliers that would be found in samples of 2000 for each group.

Whether matching is on the same or a different variable does not seem very important when groups are of 500 in size. There was agreement on 12 items for the same-variable matched groups, and 11 for the different-variable matched groups, and disagreement on 10 items for same-variable matched groups and on 16 items for the different-variable matched groups. By contrast, for the N=100 groups, the number being outliers in both groups was not very different for same-variable vs. different-variable matching (7 vs. 5), but the number of item outliers in one group but not in the other was quite high and very different, 14 in same-variable matching but 25 in different-variable matching. It seems clear that using small (N=100) samples and matching on a different variable will result in different items being considered outliers than would be considered outliers using large (N=2000) samples matched on the same variable.

CONCLUSIONS

The purpose of the study was to determine whether small sample sizes yielded the same or similar results as large, whether matching could be simplified, and whether matching on a different variable would give the same results. No problems were noticed with matching using the SPSS procedure. However, using smaller samples results in different items being identified as outliers, to some extent. The extent is more severe as the sample size is decreased. The impression is that sample sizes of N=100 for each group are intolerable, but sizes of N=500 in each group might be satisfactory for identifying the items most likely to be potentially biased. Matching on a different variable also results in somewhat different items being outliers, but to a greater extent when the sample sizes are small (100) than moderate (500).

ADDENDUM

While the study being reported is concerned only with the identification of potentially biased items, the results of the second part of the original study in which we asked judges to evaluate these items may be of interest.

The judges who were to determine whether the discrepant items were biased were chosen in terms of several criteria. First of all, they had to be people who had not been involved in writing items or specifications for items for these tests. Second, they had to have school experience. Third, they had to be people who were not known to be ardent opponents of testing in general. Fourth, each of the ethnic groups under study, white, black, Hispanic, as well as male and female, had to be represented by at least three judges. Finally, some representation of various parts of the State of Florida and of rural and urban locations was sought. Eleven judges were used.

The judges met for one day in Tallahassee to evaluate the items. Each judge was supplied with a copy of the SSAT I and the SSAT II test from which the item data resulted. Each judge also had a list of the item numbers of the items to be evaluated. For those white and Hispanic judges who were to evaluate items in terms of favoritism to either of those groups, the items numbered on the list included discrepant items from the white-Hispanic plot for a test, neutral items from that plot, discrepant items from the male-female plot, and neutral items from that plot. The analogous set of items was listed for the judges assigned to make the white-black judgments. All judges were instructed to evaluate items both for the ethnic comparison to which they were assigned and for the male-female comparison.

The results from the judges' evaluations were analyzed by recording for each item whether each appropriate judge had indicated that the item was biased and for which group. In the ethnic comparisons, items with three or more judges out of six agreeing on the fact that an item was biased and agreeing on the group favored by the item, with no other judge suggesting that the item was biased in favor of the other group, were then examined in terms of whether the judges agreed on the reason for the bias. For the comparison between males and females, there were 11 judges. If 5 or more agreed on the existence of bias and the direction of bias, and if no other judge indicated bias in the opposite direction, an item was analyzed further.

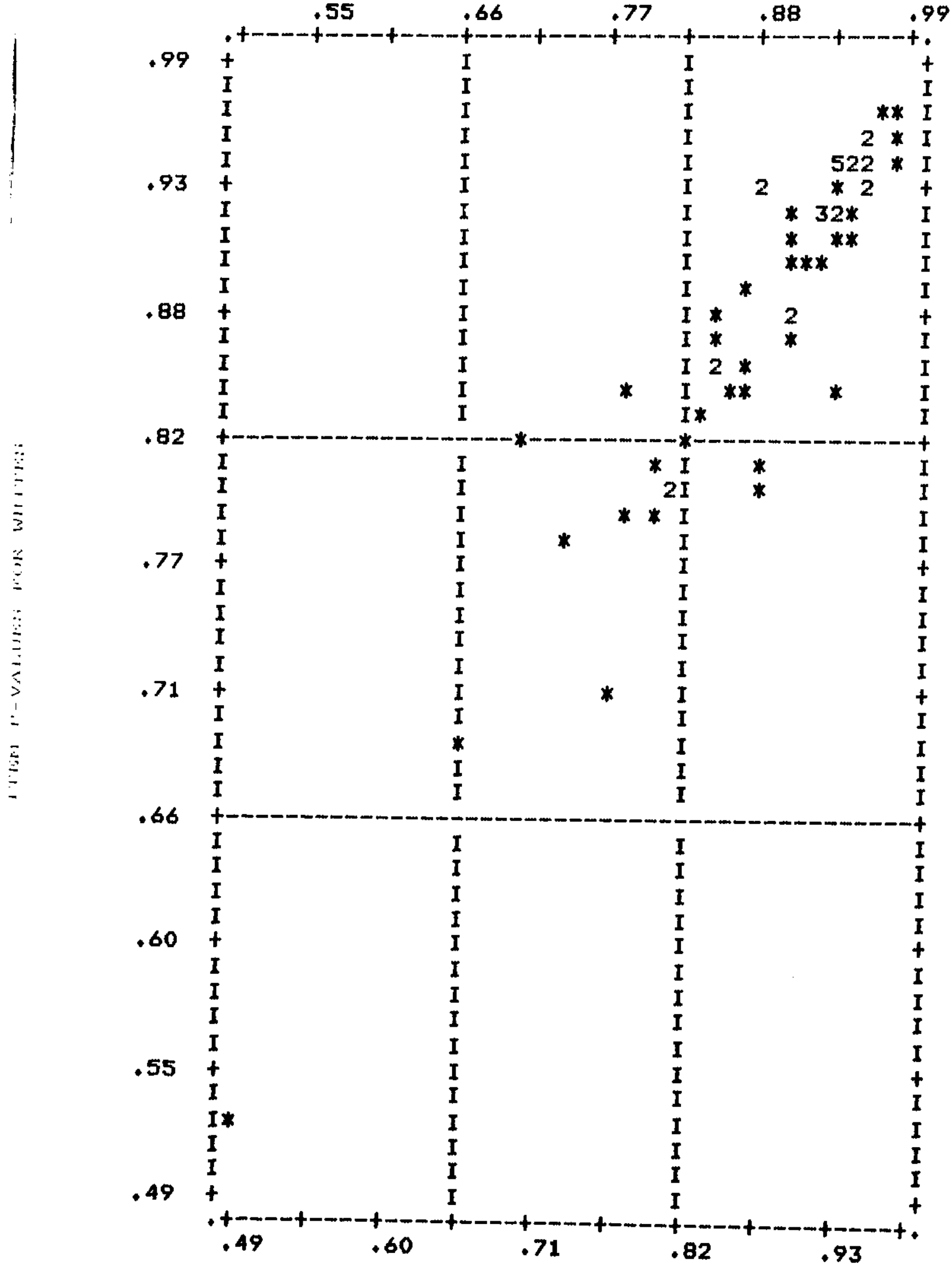
Agreement occurred on very few of the 82 items selected in the plots as discrepant. For all of the mathematics items, judges agreed on only one in terms of its favoring a particular group. The agreement between the judges' evaluations and the plots was checked to determine whether the judges consistently chose the discrepant items instead of the neutral items as indicated in the plots, whether they labeled the direction of bias correctly, and whether they agreed on the reasons they gave for the bias.

Out of all 82 items picked as discrepant in the plots of item difficulties, judges identified only 4 as favoring the same group identified in the empirical data. They failed to identify nearby items (in the plots) as discrepant in two of these cases. They identified as discrepant five items that were clearly not discrepant. Twice they identified items as discrepant but got the direction of favoritism wrong. To say the least, this is an inconclusive outcome. Apparently whatever is causing some items to be dislocated from the axis of the plot of $\underline{p}$ values is not readily discerned by expert judges.

One might be slightly encouraged in spite of the above data if when the judges did correctly choose an item as favoring one group or another they also agreed on the reason for the favoritism. The results of an analysis of their agreement on reasons for favoritism, however, are so mixed that they give no encouragement at all. Just because judges agree on the direction of favoritism is no assurance that an item is indeed discrepant or is discrepant in the direction chosen by the judges.

APPENDIX A
EXAMPLES OF P VALUE PLOTS BY GROUP

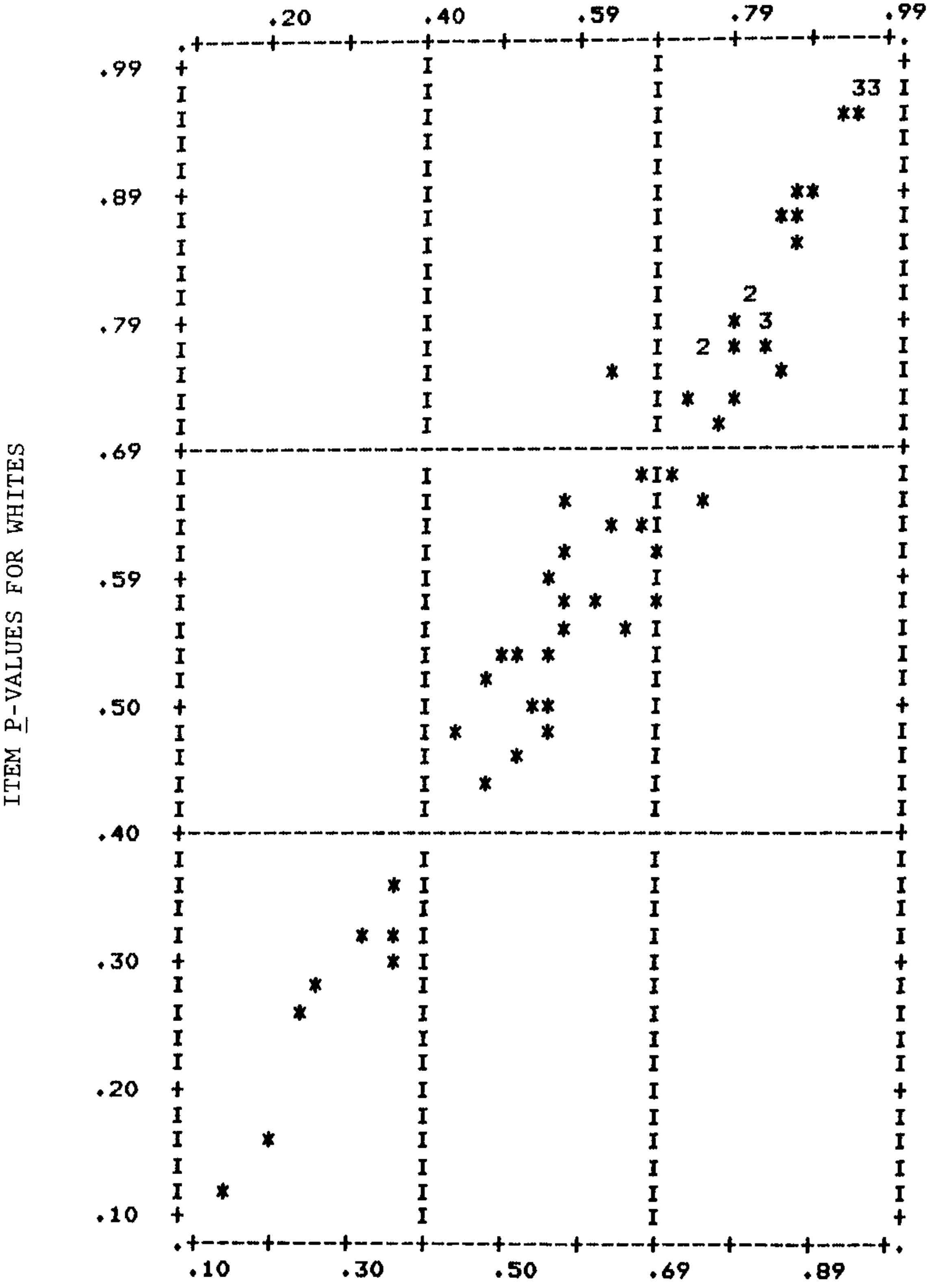
APPENDIX FIGURE 1



ITEM P - VALUES FOR BLACKS

SSAT II, Communications Items, Groups Matched on Communications Achievement

APPENDIX FIGURE 2



ITEM P - VALUES FOR BLACKS

SSAT II, Mathematics Items, Groups Matched on Mathematics Achievement

REFERENCES

Hills, J.R., and King, F.J. Detecting biased items in minimal competency tests. Paper presented to Fifth International Symosium on Educational Testing, Stirling, Scotland, June 29, 1982.

Hunter, J.E. A critical analysis of the use of item means and item test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at NIE Conference on Test Bias, Annapolis, MD., December 2-5, 1975.

Nie, H.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K. and Bent, D.H. Statistical Package for the Social Sciences. Second Edition. New York: McGraw-Hill, 1975.

Reckase, M.D. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 1979, 4, 207-230.