R E P O R T R E S U M E S

ED 011 269 UD 001 605

AN INVESTIGATION OF ITEM BIAS.

BY- CLEARY, T. ANNE HILTON, THOMAS L.
EDUCATIONAL TESTING SERVICE, PRINCETON, N.J.

REPORT NUMBER RDR-65-6-NO-12 PUB DATE APR 66
REPORT NUMBER RB-66-17
EDRS PRICE MF-$0.09 HC-$1.08 27P.

DESCRIPTORS- *APTITUDE TESTS, ACADEMIC APTITUDE, *SOCIOECONOMIC STATUS, *RACIAL DIFFERENCES, *STATISTICAL ANALYSIS, PRINCETON

THE PURPOSE OF THIS INVESTIGATION WAS TO DETERMINE WHETHER THE PRELIMINARY SCHOLASTIC APTITUDE TEST PRESENTED A DIFFERENTIAL DIFFICULTY FOR RACIAL AND SOCIOECONOMIC GROUPS. THE SUBJECTS WERE TWO GROUPS TOTALING 1,410 NEGRO AND WHITE HIGH SCHOOL SENIORS IN AN INTEGRATED HIGH SCHOOL WHO HAD TAKEN THE TEST. THEY WERE DIVIDED INTO THREE SOCIOECONOMIC LEVELS ON THE BASES OF FATHER'S OCCUPATION, FATHER'S AND MOTHER'S EDUCATION, AND A SPECIAL INDEX (HOUSE-HOME). A THREE-FACTOR ANALYSIS OF VARIANCE DESIGN (RACE, SOCIOECONOMIC STATUS, AND ITEMS ON THE MATHEMATICAL AND VERBAL SECTIONS OF THE EXAMINATION) WAS USED TO INTERPRET THE RESULTS. THE AUTHORS FOUND THAT FEW ITEMS PRODUCED AN UNCOMMON DISCREPANCY BETWEEN THE PERFORMANCE OF NEGRO AND WHITE STUDENTS AND THAT, IF THE TEST SCORES WERE DISCRIMINATORY, THE DISCRIMINATION WAS A RESULT OF PARTICULAR ITEMS ON THE TEST RATHER THAN OF THE TEST AS A WHOLE. (NH)

COLLEGE ENTRANCE EXAMINATION BOARD
RESEARCH AND DEVELOPMENT REPORTS

RDR-65-6, No. 12

RESEARCH BULLETIN
RB-66-17 APRIL 1966

An Investigation of Item Bias

T. Anne Cleary
and
Thomas L. Hilton
Developmental Research Division

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY
BERKELEY, CALIFORNIA

College Entrance Examination Board

RESEARCH AND DEVELOPMENT REPORTS

RDR-65-6, No. 12

AN INVESTIGATION OF ITEM BIAS

T. ANNE CLEARY, Developmental Research Division
and
THOMAS L. HILTON, Developmental Research Division

Research Bulletin

RB-66-17

April 1966

EDUCATIONAL TESTING SERVICE

Princeton, New Jersey

An Investigation of Item Bias

Abstract

For this research, bias was defined as an item x group interaction. Grade 12 students in integrated high schools who had taken two forms of the PSAT were divided into two races and three socioeconomic levels within the races. Four analyses of variance were performed: one for the Verbal section and one for the Mathematical section of each of the two forms of the PSAT. Because of the large sample sizes used, the tested effects were expected to be and were significant. However, only a minimal percentage of the total variance was contributed by the item x group interactions. On this basis, it was concluded that, if PSAT scores are discriminatory, the discrimination is not largely attributable to particular items, but to the test as a whole.

Table of Contents

An Investigation of Item Bias

Abstract

For this research, bias was defined as an item x group interaction. Grade 12 students in integrated high schools who had taken two forms of the PSAT were divided into two races and three socioeconomic levels within the races. Four analyses of variance were performed: one for the Verbal section and one for the Mathematical section of each of the two forms of the PSAT. Because of the large sample sizes used, the tested effects were expected to be and were significant. However, only a minimal percentage of the total variance was contributed by the item x group interactions. On this basis, it was concluded that, if PSAT scores are discriminatory, the discrimination is not largely attributable to particular items, but to the test as a whole.

An Investigation of Item Bias[1]

T. Anne Cleary and Thomas L. Hilton

As the scope of educational testing has increased, there has been a omitant increase in concern about the applicability of widely used tests ifferent cultural groups. The College Entrance Examination Board, for ple, has become concerned about the appropriateness of the Scholastic tude Test (SAT) and the Preliminary Scholastic Aptitude Test (PSAT) for subgroups of the population, particularly Negro Americans.

It is often difficult to determine what is meant by the word "bias" when s used in reference to tests. Test "bias" is explored here in terms of vidual test items. An item of a test is said to be biased for members of rticular group if, on that item, the members of the group obtain an age score which differs from the average score of other groups by more ess than expected from performance on other items of the same test. is, the biased item produces an uncommon discrepancy between the ormance of members of the group and members of other groups. In terms he analysis of variance, bias is defined as an item x group interaction. e can be no connotation of "unfair" associated with this definition of . The mean of the particular group may be higher or lower than expected.

Previous research has indicated that there are few, if any, sets of ns in the SAT which show unusual discrepancies between the performance of ro and white students. Roberts (1962) did an item analysis of a 1961 of the SAT administered to a sample of Fisk University freshmen and pared his results with those of a College Board national sample. For the analysis, Roberts used the upper and lower 5% cases from a sample of

5 students. On the average, the SAT items were more difficult for Fisk udents than for the national sample, but there was no evidence of bias in rticular items. The deltas[2] for the Fisk sample tended to be about three ints higher than those for the College Board sample, but the variance of e differences was approximately what had been observed in other nonrandom mples. The Roberts study also indicated that timing was not an important ctor in lowering the Fisk scores: The later items in the test did not differ ticeably from the earlier items in degree of discrimination or difficulty.

Cardall and Coffman (1964) applied the analysis of variance design for o-factor experiments with repeated measures on one factor to the problem item bias. In their suggested use of the design, several random samples e drawn from each of the groups being compared in order to allow the riance within groups to be used as an estimate of error. This design lows the testing of two hypotheses which are of interest in the study of em bias. First, are there significant group main effects, that is, do e groups differ significantly in mean scores? Second, is there a signifi- nt interaction between items and groups, that is, are selected items latively easier for one group than for another? If there is no signifi- nt interaction, one may conclude that the test is homogeneous across groups, d that, if a difference in item difficulty exists, it is present equally all items.

From the May 1963 administration of the SAT, Cardall and Coffman drew ree samples of 300 cases from each of three groups: Group 1 answer sheets re selected from rural centers in the midwest, group 2 answer sheets were lected from centers in New York City, and group 3 answer sheets were lected from centers in the southeast where only Negro candidates were gistered.

Two analyses of variance were performed, one for the 40 verbal items in Section II of the test, and the other for 25 mathematical items in Section III. For both the verbal and mathematical items, Cardall and Coffman found significant group main effects. The major differences in both cases were between groups 1 and 3 and groups 2 and 3: the mean performance of group 3 was lower in each case.

The interaction between groups and items was also highly significant in both analyses. Since there were three samples in each of the three groups, Cardall and Coffman were able to compute independent correlations of item difficulties within groups and between groups. Two of the samples within each group were used to find the within-group correlations between item difficulties; the third sample in each group was used to find the correlations between the item difficulties of the different groups. The within-group correlations indicated the degree to which the item difficulties varied in different samples from the same group. The between-group correlations indicated the degree to which the relative difficulty of the items remained constant from one group to the next.

For both verbal and mathematical items, the within-group correlations were very high (between .96 and .99). For the verbal items, the between-group correlations involving group 3 were much lower than the within-group correlations. Thus, a major factor in determining the significance of the interaction for verbal items appeared to be the lack of correspondence between the relative difficulties of the items for group 3 and the other two groups. Since the three between-group correlations for the mathematical items were similar, it appeared that the factors accounting for the significant interaction were evenly distributed across the three groups.

The Cardall and Coffman analysis indicated that the SAT items did not retain the same relative difficulty across groups, but the analysis did not indicate whether the discrepancies were all in one direction or balanced so that one group was not favored over another. Fremer[3] continued the Cardall and Coffman study by plotting the arcsin transformations of the item difficulties for each pair of groups. Fremer found that two items with a distinctly rural flavor were very much easier for group 1 than for the other two groups. Thus it might be said that the test has a slight rural bias. When group 3 was compared with each of the other groups, a slight curvilinearity was found in the plots, but this appeared to be due to a "floor" effect in the group 3 responses. This curvilinearity would have attenuated the correlations between the item difficulties of group 3 and the other groups.

Purpose

The purpose of this research was to study the variation of Preliminary Scholastic Aptitude Test (PSAT) item scores in different racial and socio-economic (SES) groups. The questions asked were whether the test items are equally difficult for all groups, whether the group mean scores across items differ by groups, or whether both group means and relative scores on individual items change as a function of race, SES within race, or both. Although the primary question at hand was the possibility of differential difficulty for racial groups, the inclusion of SES as a factor in the research made it possible to study this variation as well as the variation associated with race.

Procedure

Sample. Every two years, as a part of a longitudinal study of academic growth, a large sample of twelfth-grade students is given the PSAT. For this research, seven integrated schools in three large metropolitan centers were selected from the larger sample, and the race of the 1961 (Group I) and 1963 (Group II) twelfth-graders was identified by the school administrators. In order to have an equal number of students of each race, it was necessary to use all available Negro students and randomly sample the white students. For the analyses, Group I had 636 students; Group II, 774.

Variables. Group I took form IPT1 of the PSAT; Group II took form KPT (a parallel form). The five-option items were scored by giving one point for a correct response, zero for no response, and minus one-quarter for an incorrect response. This scoring method was based on the formula used to obtain the total test score.

SES was defined by information from a background and experience questionnaire which was completed by Groups I and II in 1963. An SES score was obtained from questions on father's occupation, mother's and father's education, and the House-Home Index (Kerr & Remmers, 1942). Students within each race were then ordered from high to low and divided into three equal SES groups: high, middle, and low SES. In all cases, the cutting scores for the SES levels were lower for the Negro SES groups than for the white groups.

Method of analysis. A three-factor analysis of variance design was used. Figure 1 gives a schematic representation of the design. The first factor was race, which was considered a fixed factor. A second factor was

SES, which was considered fixed and nested within race. The nesting of SES within race avoided the assumption that the SES levels are comparable in the two races. A third factor was items, which was considered random.

The linear model for the design had the form:

$$E(X_{pirs}) = \mu + R_r + S_{s(r)} + I_i + P_{p(rs)} + IS_{is(r)} + IR_{ir} + IP_{ip(rs)}$$

where

X_{pirs} = the score of the p^{th} person of the r^{th} race and s^{th} socioeconomic level on the i^{th} item,

μ = the grand mean,

R_r = effect of the r^{th} race, $r = 1, \ldots, N_r$,

$S_{s(r)}$ = effect of the s^{th} socioeconomic level within the r^{th} race, $s = 1, \ldots, N_s$,

I_i = effect of the i^{th} item, $i = 1, \ldots, N_i$,

$P_{p(rs)}$ = effect of the p^{th} person, $p = 1, \ldots, N_p$ within the r^{th} race and s^{th} socioeconomic level,

$IS_{is(rs)}$ = interaction of items and socioeconomic level within race,

IR_{ir} = interaction of items and races,

$IP_{ip(rs)}$ = interaction of items and persons within race and socioeconomic level.

Table 1 gives the summary of the analysis of variance design. The expected mean squares were derived by the method of Cornfield and Tukey (1956). In this summary, race and socioeconomic level are considered fixed effects, the items and persons random.

Two analyses were performed for each group: one for the 70 items of the verbal section of the PSAT, and one for the 50 items of the mathematical

section. The analysis of variance design allowed the testing of five hypotheses:

(1) $\sigma^2_{IR} = 0$. That there is no interaction between items and race. This is the most important hypothesis. As bias has been defined here for items, an interaction between items and race would indicate presence of racial bias, and that the pattern of item difficulties is different in the two racial groups.

(2) $\sigma^2_{IS(R)} = 0$. That there is no interaction between items and SES within race. The presence of item-SES interaction would indicate that the items are biased for different SES groups within at least one of the two races.

(3) $\sigma^2_{I} = 0$. That there is no difference in the mean scores of different items. It was expected that this hypothesis would be rejected at a very high level of significance because it was known that the items do differ considerably in difficulty.

(4) $\sigma^2_{S(R)} = 0$. That there is no difference in mean item scores in different SES levels within race. It was expected that this hypothesis would be rejected, and that the higher SES levels would have higher mean scores. This hypothesis must be tested by a quasi F ratio (Satterthwaite, 1946).

(5) $\sigma^2_{R} = 0$. That there is no difference in mean item scores for the two races. This hypothesis must be tested by a quasi F ratio.

Results

Tables 2 and 3 contain summaries of the analyses of variance for Groups I and II. With 106 and 129 persons in each of the six Race x SES groups,

even rather small, inconsequential differences may be significant. As was expected, almost all tested effects were found to be significant.

A more meaningful way of looking at the results of the analyses of variance with such large sample sizes is to estimate the variance components and the percentage contribution of each effect to the total variance of a single observation chosen at random. These figures are given in Table 4. It should be remembered that a single observation is an item score which may have the values one, zero, or minus one-quarter. In all four analyses, a major percentage of the variance was contributed by the Subject x Item interaction. The Subject x Item interaction is treated by Hoyt (1941) as the variance due to error of measurement. By subtracting from one the proportion of variance due to Subject x Item interaction, an estimate of the reliability of an item can be obtained. Large percentages of the total variance were contributed by the effect of Persons Within Race-SES Groups and the effect of Items. The smallest contributions to the total variance were provided by the Item x Race interaction (the indicator of racial bias) and Item x SES Within Race interaction (the indicator of social class bias). Given the stated definition of bias (an item x group interaction), the PSAT cannot for practical purposes be considered biased for either race or SES within race.

The lack of a practically significant amount of item x group interaction is made clear in Figures 2 through 5 which contain bivariate plots of the sums of item scores for Negro and white students. From these plots, it can be seen that the items are on the average easier for the white students: the concentration of points is below the 45 degree line. However, there

appears to be no systematic deviation from a straight line except possibly that caused by what appears to be "floor" effect in the Negro scores. The "floor" effect, indicated by the decreased slope at the left of the plots, would contribute to the item x race interaction.

Discussion

For this research, bias was defined as an item x group interaction. In four separate analyses, the Item x Race and Item x SES Within Race interactions contributed minimal percentages of the total variance of an observation. From the bivariate plots of sums of item scores, it was apparent that there were few items producing an uncommon discrepancy between the performance of Negro and white students. It must therefore be concluded that, given the stated definition of bias, the PSAT for practical purposes is not biased for the groups studied. The question of bias as a total test score difference between groups has not been considered here. A second phase of this research, now in progress, is designed to investigate total test score differences and the way in which these differences affect the predictive validity of the SAT.

References

Cardall, Carolyn, & Coffman, W. E. A method for comparing the performance of different groups on the items in a test. Research Bulletin 64-61, Princeton, N. J.: Educational Testing Service, 1964.

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27, 907-949.

Hoyt, C. Test reliability obtained by analysis of variance. Psychometrika, 1941, 6, 153-160.

Kerr, W. A., & Remmers, H. H. Manual for the American Home Scale. Chicago, Ill.: Science Research Associates, 1942.

Roberts, S. O. Studies in identification of college potential. Mimeographed Report, Fisk University, 1962.

Satterthwaite, F. E. An approximate distribution of estimates of variance components. Biometrics Bulletin, 1946, 2, 110-114.

Footnotes

[1]The authors are grateful to William H. Angoff, William E. Coffman, and Julian C. Stanley for numerous suggestions and helpful criticisms.

[2]Delta is defined as the normal deviate, expressed in terms of a scale with mean of 13 and standard deviation of 4, which corresponds to the proportion of candidates reaching the item who answer it correctly. A low delta describes an easy item; a high delta, a difficult one.

[3]Fremer, J. Manuscript in preparation, 1966.

Table 1

Summary of the Analysis of Variance[a]

Source of Variation	df	E(MS)
Between Subjects	$N_p N_r N_s - 1$	
F_r (Race)	$N_r - 1$	$\sigma^2_\epsilon + \sigma^2_{IP(RS)} + N_s N_p \sigma^2_{IR} + N_i \sigma^2_{P(RS)} + N_s N_i N_p \sigma^2_R$
$F_{s(r)}$ (SES Within Race)	$N_r(N_s - 1)$	$\sigma^2_\epsilon + \sigma^2_{IP(RS)} + N_p \sigma^2_{IS(R)} + N_i \sigma^2_{P(RS)} + N_i N_p \sigma^2_{S(r)}$
$F_{p(RS)}$ (Persons Within Race and SES)	$N_r N_s (N_p - 1)$	$\sigma^2_\epsilon + \sigma^2_{IP(RS)} + N_i \sigma^2_{P(RS)}$
Within Subjects	$N_p N_r N_s (N_i - 1)$	
F_i (Items)	$N_i - 1$	$\sigma^2_\epsilon + \sigma^2_{IP(RS)} + N_r N_s N_p \sigma^2_I$
F_{ir} (Items x Race)	$(N_i - 1)(N_r - 1)$	$\sigma^2_\epsilon + \sigma^2_{IP(RS)} + N_s N_p \sigma^2_{IR}$
$F_{is(r)}$ (Items x SES Within Race)	$N_r(N_s - 1)(N_i - 1)$	$\sigma^2_\epsilon + \sigma^2_{IP(RS)} + N_p \sigma^2_{IS(R)}$
$F_{ip(rs)}$ (Items x Persons Within Race and SES)	$N_r N_s (N_i - 1)(N_p - 1)$	$\sigma^2_\epsilon + \sigma^2_{IP(RS)}$

[a] R (race) and S (socioeconomic level) are considered fixed effects; I (items) and P (persons) are considered random effects.

Table 2

Summary of the Group I Analyses of Variance

PSAT Verbal

Source	df	SS	MS	F
Between Subjects	635	1,608.37		
Race	1	208.64	208.64	103.3*
SES Within Race	4	131.09	32.77	15.7*
Between Subjects Nested Within Race-SES Groups	630	1,268.64	2.01	
Within Subjects	43,884	13,813.62		
Items	69	2,030.76	29.43	110.3
Items x Race	69	89.53	1.30	4.9
Items x SES Within Race	276	95.75	.35	1.3
Interaction of Subjects with Items Within Race-SES Groups	43,470	11,597.58	.27	
Total	44,519	15,421.99		

PSAT Math

Source	df	SS	MS	F
Between Subjects	635	1,818.69		
Race	1	296.77	296.77	132.2*
SES Within Race	4	115.16	28.79	12.9*
Between Subjects Nested Within Race-SES Groups	630	1,406.76	2.23	
Within Subjects	31,164	8,433.96		
Items	49	990.40	20.21	85.5
Items x Race	49	102.97	2.10	8.8
Items x SES Within Race	196	47.04	.24	1.0
Interaction of Subjects with Items Within Race-SES Groups	30,870	7,293.55	.24	
Total	31,799	10,252.65		

*Quasi F ratios

Table 3

Summary of the Group II Analysis of Variance

PSAT Verbal

Source	df	SS	MS	F
Between Subjects	773	1,831.48		
Race	1	382.61	382.61	215.0*
SES Within Race	4	91.19	22.80	12.8*
Between Subjects Nested Within Race-SES Groups	768	1,357.68	1.77	
Within Subjects	53,406	16,859.17		
Items	69	2,334.69	33.84	130.2
Items x Race	69	164.47	2.38	9.2
Items x SES Within Race	276	94.88	.34	1.3
Interaction of Subjects with Items Within Race-SES Groups	52,992	14,265.13	.26	
Total	54,179	18,690.65		

PSAT Math

Source	df	SS	MS	F
Between Subjects	773	1,431.31		
Race	1	221.72	221.72	147.2*
SES Within Race	4	68.33	17.08	11.4*
Between Subjects Nested Within Race-SES Groups	768	1,141.26	1.49	
Within Subjects	37,926	10,999.68		
Items	49	2,082.63	42.50	184.8
Items x Race	49	125.26	2.55	11.1
Items x SES Within Race	196	61.89	.32	1.4
Interaction of Subjects with Items Within Race-SES Groups	37,632	8,729.90	.23	
Total	38,699	12,430.99		

*Quasi F ratios

Table 4

Estimated Components of Variance

Group I

Effect	Verbal Est. σ^2	Verbal % of Total	Math Est. σ^2	Math % of Total
Race	.00923	2.6	.01841	5.5
SES Within Race	.00413	1.2	.00501	1.5
Persons Within Race-SES Groups	.02495	7.0	.03993	11.9
Items	.04586	12.9	.03141	9.3
Items x Race	.00324	.9	.00587	1.7
Items x SES Within Race	.00076	.2	.00004	0.0
Interaction of Subjects with Items Within Race-SES Groups	.26680	75.2	.23637	70.1

Group II

Effect	Verbal Est. σ^2	Verbal % of Total	Math Est. σ^2	Math % of Total
Race	.01398	3.9	.01126	3.4
SES Within Race	.00232	.7	.00240	.7
Persons Within Race-SES Groups	.02141	6.0	.02508	7.6
Items	.04337	12.2	.05461	16.4
Items x Race	.00546	1.5	.00601	1.8
Items x SES Within Race	.00058	.2	.00065	.2
Interaction of Subjects with Items Within Race-SES Groups	.26919	75.5	.23198	69.9

		Items I_1 I_2 ... I_{N_1}
R_1 Negro	$S_{3(1)}$ High SES	Person 1 (3,1) Person 2 (3,1) . . . . . . Person N_p (3,1)
	$S_{2(1)}$ Middle SES	Person 1 (2,1) Person 2 (2,1) . . . . . . Person N_p (2,1)
	$S_{1(1)}$ Low SES	Person 1 (1,1) Person 2 (1,1) . . . . . . Person N_p (1,1)
R_2 White	$S_{3(2)}$ High SES	Person 1 (3,2) Person 2 (3,2) . . . . . . Person N_p (3,2)
	$S_{2(2)}$ Middle SES	Person 1 (2,2) Person 2 (2,2) . . . . . . Person N_p (2,2)
	$S_{1(2)}$ Low SES	Person 1 (1,2) Person 2 (1,2) . . . . . . Person N_p (1,2)

Figure 1. Schematic representation of the analysis of variance design.

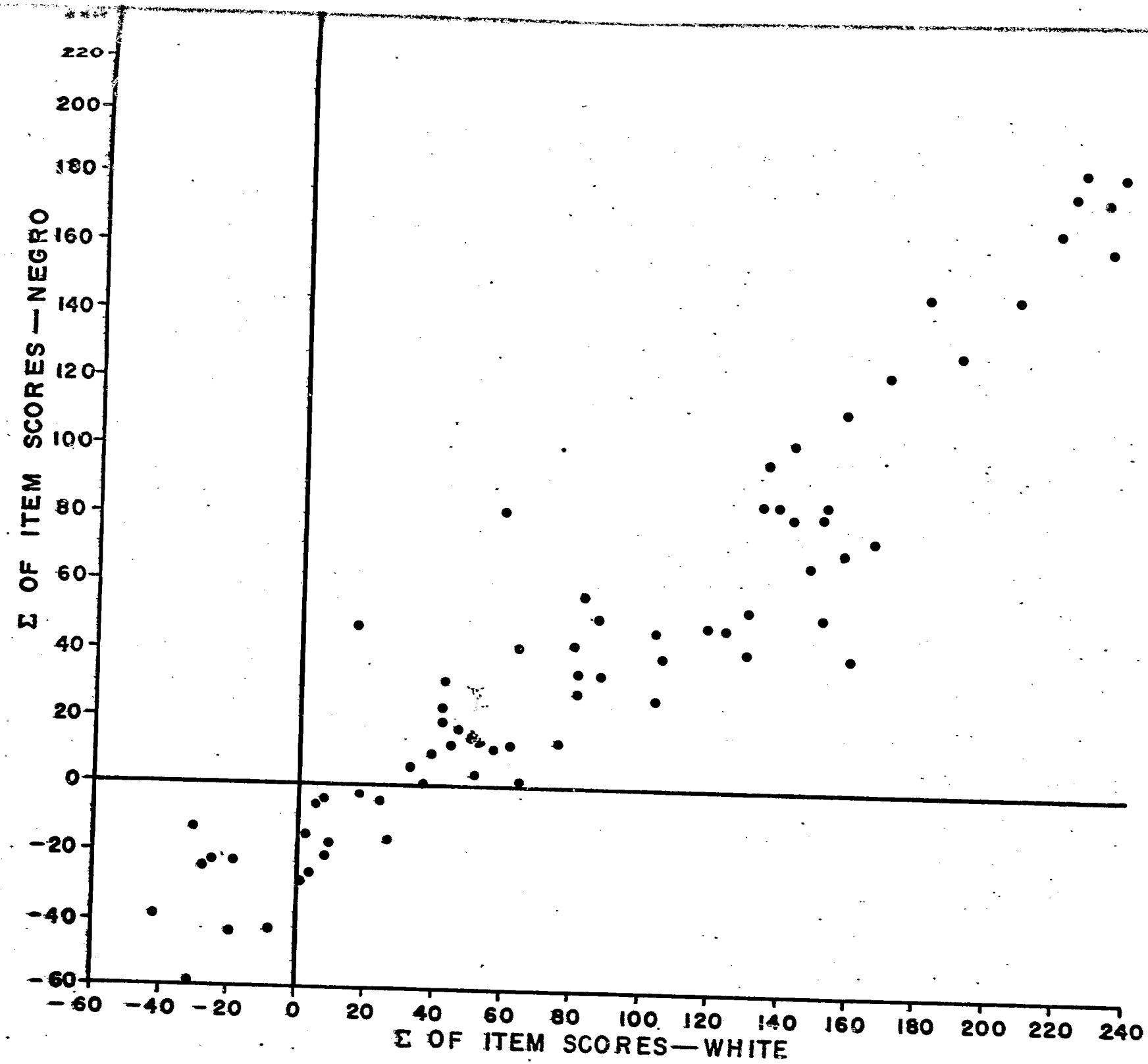

Figure 2. Bivariate Plot of Sums for PSAT Verbal Items--Group I
Sample Size for Each Race: 318

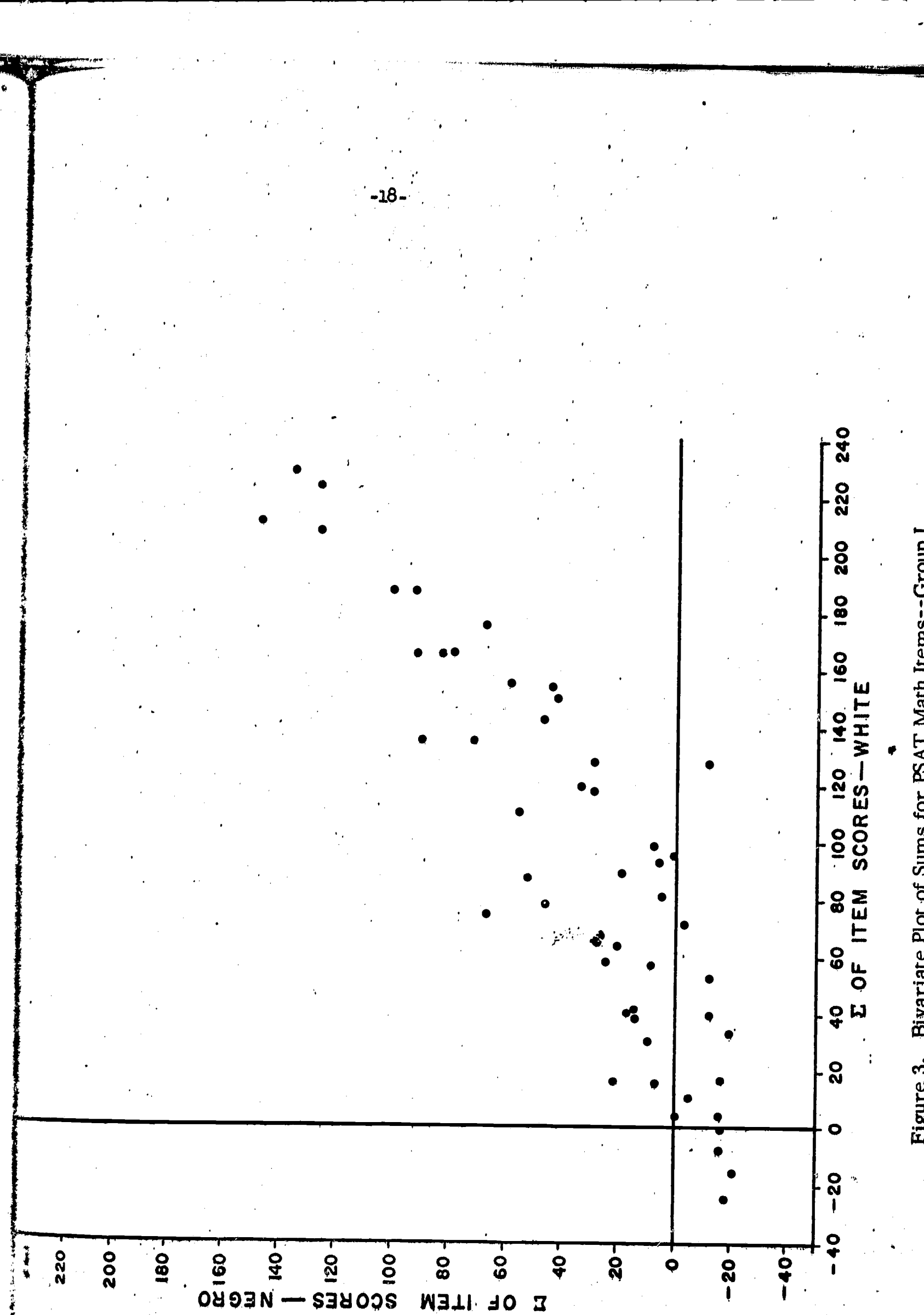

Figure 3. Bivariate Plot of Sums for PSAT Math Items--Group I
Sample Size for Each Race: 318

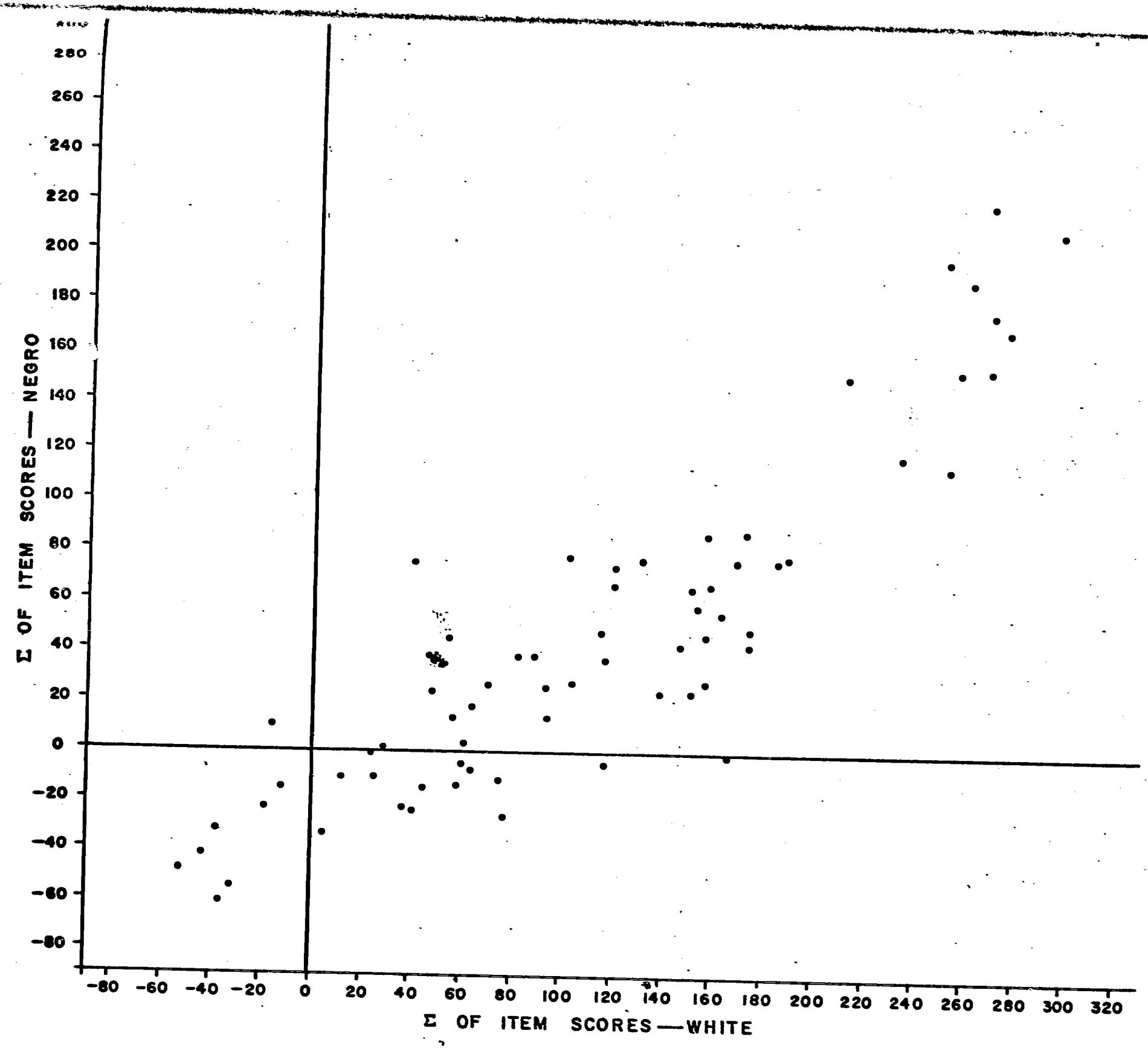

Figure 4. Bivariate Plot of Sums for PSAT Verbal Items--Group II
Sample Size for Each Race: 387

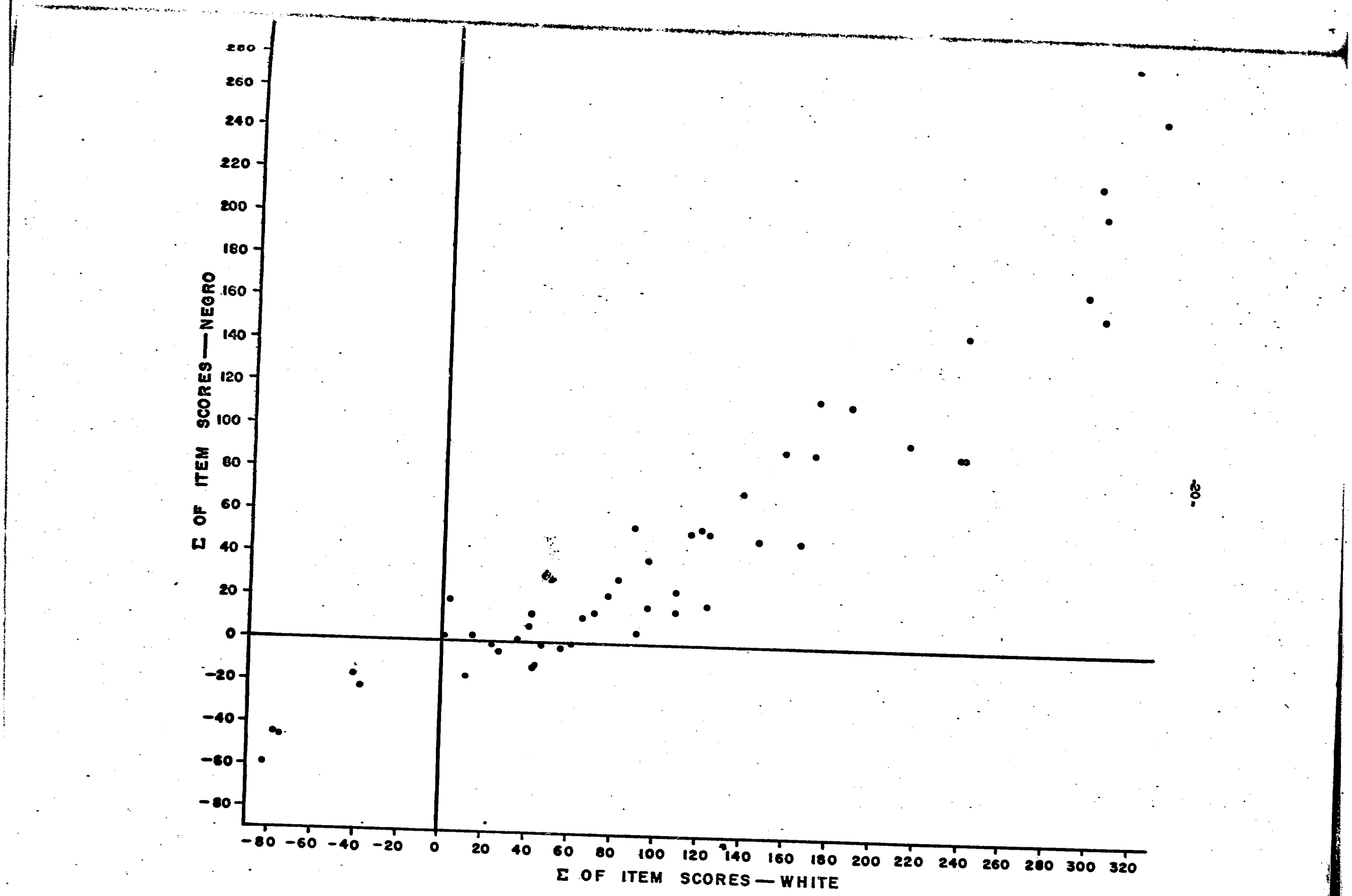

Figure 5. Bivariate Plot of Sums for PSAT Math Items--Group II
Sample Size for Each Race: 387