DOCUMENT RESUME

ED 069 700 TM 002 151

AUTHOR Livingston, Samuel A.
TITLE Verbal Overload in Achievement Tests.
INSTITUTION Johns Hopkins Univ., Baltimore, Md. Center for the Study of Social Organization of Schools.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE Jun 72
GRANT OEG-2-7-061610-0207
NOTE 26p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Achievement Tests; Factor Analysis; Grade 8; Junior High School Students; Measurement Techniques; Multiple Choice Tests; Performance Tests; *Reading Ability; Reading Level; Research Methodology; *Social Studies; Statistical Analysis; Technical Reports; *Test Construction
IDENTIFIERS Dale Chall Formula for Reading Difficulty; SCAT; School and College Ability Tests; Sequential Tests of Educational Progress; STEP

ABSTRACT

A social studies achievement test made up of items rewritten in simplified language was compared with a test containing the same items in their original form by administering the two tests to the entire 8th grade class of a suburban junior high school near Baltimore. The results showed only slightly higher scores for students taking the simplified test. Differences among the items in estimated reading difficulty were not associated with differences in actual response difficulty. The findings were interpreted to mean that most students who know enough to answer a test item can also read well enough to understand it. (Author)

Center for Social Organization of Schools

REPORT No. 133

JUNE, 1972

VERBAL OVERLOAD IN ACHIEVEMENT TESTS

SAMUEL A. LIVINGSTON

STAFF

John L. Holland, Director

James M. McPartland, Assistant Director

Virginia B. Bailey

Joan E. Brown

Zahava D. Blum

Judith P. Clark

James S. Coleman

David L. DeVries

Keith J. Edwards

Gail M. Fennessey

James J. Fennessey

Linda P. Ferraris

Stephanie G. Freeman

Ellen Greenberger

Rubie J. Harris

Edward J. Harsch

Samuel T. Helms

Robert T. Hogan

John H. Hollifield

Karen L. Jaworski

Nancy L. Karweit

Steven J. Kidder

Samuel A. Livingston

Edward L. McDill

Rebecca J. Muraro

Alice J. Nafziger

Dean H. Nafziger

Julian C. Stanley

B. James Starr

John P. Snyder

Diana F. Ward

VERBAL OVERLOAD IN ACHIEVEMENT TESTS

Grant No. OEG-2-7-061610-0207

Program No. R16J1

Samuel A. Livingston

Report No. 133

June, 1972

Published by the Center for Social Organization of Schools, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the Office of Education, and no official endorsement by the Office of Education should be inferred.

The Johns Hopkins University

Baltimore, Maryland

INTRODUCTORY STATEMENT

The Center for Social Organization of Schools has two primary objectives: to develop a scientific knowledge of how schools affect their students, and to use this knowledge to develop better school practices and organization.

The Center works through five programs to achieve its objectives. The Academic Games program has developed simulation games for use in the classroom, and is studying the processes through which games teach and evaluating the effects of games on student learning. The Social Accounts program is examining how a student's education affects his actual occupational attainment, and how education results in different vocational outcomes for blacks and whites. The Talents and Competencies program is studying the effects of educational experience on a wide range of human talents, competencies, and personal dispositions in order to formulate--and research--important educational goals other than traditional academic achievement. The School Organization program is currently concerned with the effects of student participation in social and educational decision-making, the structure of competition and cooperation, formal reward systems, ability-grouping in schools, and effects of school quality. The Careers and Curricula program bases its work upon a theory of career development. It has developed a self-administered vocational guidance device to promote vocational development and to foster satisfying curricular decisions for high school, college, and adult populations.

This report, like others occasionally published by the Center, deals with a subject common to all programs -- that of scientific measurement.

ABSTRACT

A social studies achievement test made up of items rewritten in simplified language was compared with a test containing the same items in their original form by administering the two tests to the entire 8th-grade class of a suburban junior high school near Baltimore. The results showed only slightly higher scores for students taking the simplified test. Differences among the items in estimated reading difficulty were not associated with differences in actual response difficulty. The findings were interpreted to mean that most students who know enough to answer a test item can also read well enough to understand it.

ACKNOWLEDGMENTS

I thank Harry Bornstein for his advice in planning this replication of his experiment and for providing the necessary test materials. I thank George Brauer for supervising the actual administration of the tests. Without their help, this experiment would not have been possible. I also thank David DeVries for his many helpful comments on an earlier version of this paper.

Introduction

Multiple-choice tests are commonly used to test many different kinds of knowledge and skills. The poor reader would appear to be at an obvious disadvantage when taking this type of test, and his disadvantage would seem to be greatest when the items are written in language which is more complex than it has to be. Bornstein and Chamberlain (1970, p. 597) have argued that "the language used on multiple-choice achievement test items should be no more complex than is necessary to test the examinee's knowledge of the subject matter. Language complexity above this minimum level can be regarded as verbal overload and may constitute a source of bias against those people whose verbal skills are limited."

To find out whether "verbal overload" actually affects examinees' test scores in a school testing situation, Bornstein and Chamberlain (1970) used a test made up of items from STEP[1] social studies tests. These items measure the student's ability to interpret social studies materials. The items were of two types; some were based on information presented in pictorial or graphic form, while others were based on information presented in written passages. The test was printed in two forms. The pictures, graphs, and written passages were identical on the two forms, but the wording of the items was different. One form contained the items as originally written; the other contained the same items, rewritten in simplified language and reproduced in larger type.

---

[1] Sequential Tests of Educational Progress, published by Educational Testing Service.

Bornstein and Chamberlain's subjects were junior and senior high school students in Oakland, California. Despite the students' generally low reading ability (their mean was at about the 30th percentile on national norms), the students who took the simplified form of the test failed to outperform those who took the test containing the items in their original form. Bornstein attributed this result to lack of motivation on the part of the subjects.[1] Bornstein (1971) later performed a similar study using the same materials with deaf college-preparatory students. He found small but significant differences in favor of the students who took the simplified form of the test. Bornstein and Kannapell (1971) replicated this study with a broad sample of deaf high school students and found no significant differences between the groups taking the different forms of the test.

The present study was basically a replication of Bornstein and Chamberlain's experiment, with a different subject population and with a few additional refinements in the design and analysis. Because of Bornstein's suspicions that lack of motivation on the part of his inner-city subjects may have been responsible for his finding of no difference, this replication was conducted with suburban students. Since the simplification of the items might be expected to help only the poor readers, the students' verbal ability was considered as a factor in the design. And because the rewriting of the test items seemed to simplify some items more than others, estimates were made of the reading difficulty of each item in its original and simplified versions. This experiment

[1] Personal communication, March, 1971.

can therefore be considered a test of the following three hypotheses:

1. In general, there will be more correct responses to the rewritten items than to the original items.
2. This difference will be greatest for students of low verbal ability.
3. Those items which show the greatest decrease in estimated reading difficulty when rewritten will show the greatest increase in proportion of correct responses.

Method

The materials used in the present study were the same materials used by Bornstein and Chamberlain (1970) and by Bornstein (1971).[1] The test consisted of forty-eight multiple-choice items which tested the students' ability to interpret social studies materials. Items 1 to 32 were based on information presented in charts, tables, pictures, or graphs. The remaining sixteen items were based on information contained in prose passages about a half-page long. Within each of these two subtests, half the items were taken from a junior-high-school-level test; the other half from a senior-high-school-level test. Thus, the test can be considered as a single test, or two subtests, or four sub-subtests.

---

[1] I am indebted to Harry Bornstein for making these materials available for this experiment.

Figure 1 shows two of the "graphic" items in their original form; Figure 2 shows the same two items as they appeared on the simplified form of the test.

The reading difficulty of each test item, in both original and re-written form, was estimated by the Dale-Chall formula (Dale and Chall, 1948).[1] The mean estimated reading difficulty of the original items was 7.16 (9th grade level); that of the simplified items was 5.80 (6th grade level). The standard deviation of the estimated reading difficulty scores was 1.37 for the original items and 1.15 for the simplified items. These values must be considered as a rough approximation, since the Dale-Chall formula is intended for use with reading selections much longer than a single test item.

The subjects for this study were the entire eighth-grade class of a suburban junior high school near Baltimore. Their verbal ability scores ranged from the 5th to the 96th percentile on county-wide norms, with most of the scores between the 30th and 60th percentiles. The students took the tests at the end of the school year in their regular social studies classes. The tests were administered by the regular classroom teachers, who were instructed not to answer the students' questions about the test - especially, not to tell them the meanings of unfamiliar words. The teachers reported that the students were highly motivated.

The students were assigned forms of the test (original or simplified) by random selection . Ability grouping was by

[1] This formula has been extensively validated; see Klare (1963) for a discussion of it and other readability formulas.

quartiles, based on verbal scores from the SCAT,[1] administered nineteen months previously. When the sex of the students is taken into account, the resulting design is a 2 x 4 x 2 fully-crossed factorial experiment. The number of students in each cell of the design is shown in Table 1. There is a relationship between verbal ability and sex of student - a higher proportion of the students at the lower ability levels were boys. The students were allowed forty minutes for the test. About 24 percent of the students taking the original items and about 16 percent of those taking the simplified items did not finish the test.

Results

On the basis of hypothesis 1, we would expect a substantially higher score for the students taking the simplified items than for those taking the original items. This difference would be reflected in the analysis of variance as a strong effect for test form. On the basis of hypothesis 2, we would expect a pattern of scores showing a large advantage at the low end of the ability scale for those students taking the simplified items, and only a small difference at the upper end of the ability scale. This trend would be reflected in the analysis of variance by a strong form-ability interaction effect. Neither of these hypothesized effects was reflected in the observed results.

[1] School and College Ability Tests, published by Educational Testing Service.

The mean scores for the full test are shown in Table 2 and in Figure 3. The differences associated with the difference in test forms were generally small - about one or two items on a 48-item test. The simplification of the items seems to have helped the high- and average-ability boys and the low-ability girls. Tables 3 and 4 and Figure 3 show the results for the graphic and prose items separately. None of the sets of scores shows the anticipated pattern of differences, and the patterns which do appear do not suggest any reasonably simple explanation other than sampling variability.

Table 5 presents the results of analyses of variance[1] on the total scores, the two subtests (graphic and prose items), and four sub-tests. Each column in the table represents a separate three-way analysis of variance. Although all three main effects were statistically significant for the prose items, and the graphic items showed a significant three-way interaction, only the main effect for ability accounted for a substantial portion of the variance in any of the analyses. The three-way interaction on the graphic items accounts for about two percent of the variance and reflects the tendency of the simplified items to help the average-ability boys and the lower-ability girls.

The first two columns of Table 6 present the correlations of the estimated reading difficulty of the items with their actual response difficulty, as indicated by the proportion of students missing the item. These correlations are positive for both sets of items and generally

[1]These analyses were performed by means of the computer program Multivariance (Finn, 1968), which computes a least-squares solution for unequal and disproportional cell frequencies.

larger for the original items than for the rewritten items. The third and fourth columns of Table 6 present two sets of correlations which indicate the extent to which changes in the estimated reading difficulty of the items were associated with changes in their actual response difficulty. The column labeled "change" shows a correlation of the unadjusted difference between test forms (original minus rewritten) for the two types of difficulty. The column labeled "Residuals" shows the correlation of these differences, adjusted for the difficulty of the original items.[1] Hypothesis 3 would predict substantial positive correlations in these two columns, particularly in the last column. However, the correlations of these change measures are about zero overall, and in the separate subgroups of students they are as often negative as positive. Furthermore, the subgroups in which the rewritten items seemed to help the students most, as indicated by the subgroup mean scores, were not the ones in which the test items that decreased the most in estimated reading difficulty also decreased the most in actual response difficulty.

Discussion

In interpreting the results of this experiment, it is important to remember that the test was given under moderately speeded conditions.

[1] The two variables being correlated are thus measures of the extent to which the difficulty of the item changed more or less than might be expected on the basis of its original difficulty. See Lord (1963) for a more thorough explanation.

(About twenty percent of all the students did not finish the test.) Highly speeded conditions might have produced greater differences between students' scores on the two versions of the test; totally unspeeded conditions might have eliminated the small differences which did appear. The simplification of the items seems to have had some effect on the students' speed in taking the test, since the proportion of students who did not finish was smaller by one-third in the group taking the simplified items.

In general, the results of this experiment do not support the three hypotheses it was designed to test. Although the scores on the simplified test were slightly better than those on the original test, the differences were minimal. Likewise, the interaction effects involving the difference between the original and simplified tests were either nonexistent or so small as to be of no practical significance. The group mean scores, ranging from about one-third to about two-thirds of the items correct, indicate that floor or ceiling effects cannot account for this absence of sizable differences. Finally, the items which were the most simplified, according to a readability formula, were not the ones on which students taking the simplified form of the test tended to outperform those taking the original form. Therefore, Bornstein and Chamberlain's conclusion that ". . . verbal load does not appear to be a significant factor . . ." appears to be about as true in the suburbs as in the inner city.

Why should the reading difficulty of test items have so little effect on their actual difficulty? One possible explanation for Bornstein and Chamberlain's results is a lack of motivation on the part of the

students, but that explanation would hardly account for the results of the present experiment. The subjects were middle-class, suburban students, and their teachers described them as showing a high level of motivation for the test. The most plausible explanation is simply that most students who know enough of the content being tested to be able to answer a particular test item correctly can also read well enough to understand the item.

References

Bornstein, Harry. Some effects of verbal load on achievement tests. American Annals of the Deaf, 1971, 116, 44-48.

Bornstein, Harry and Chamberlain, Karen. An investigation of the effects of "verbal load" in achievement tests. American Educational Research Journal, 1970, 7, 597-604.

Bornstein, Harry and Kannapell, Barbara. More on the effects of verbal load on achievement tests. American Annals of the Deaf, 1971, 116, 575-579.

Dale, Edgar and Chall, Jeanne S. A formula for predicting readability. Educational Research Bulletin, 1948, 27, 11-20, 37-54.

Finn, Jeremy D. Multivariance: Univariate and Multivariate Analysis of Variance, Covariance, and Regression (version 4). State University of New York at Buffalo, Department of Educational Psychology, 1968.

Klare, George R. The Measurement of Readability. Ames, Iowa: Iowa State University Press, 1963.

Lord, Frederic M. Elementary models for measuring change. In Harris, Chester W. (ed.). Problems in Measuring Change. Milwaukee, University of Wisconsin Press, 1962.

Figure 1. Sample items in original form.

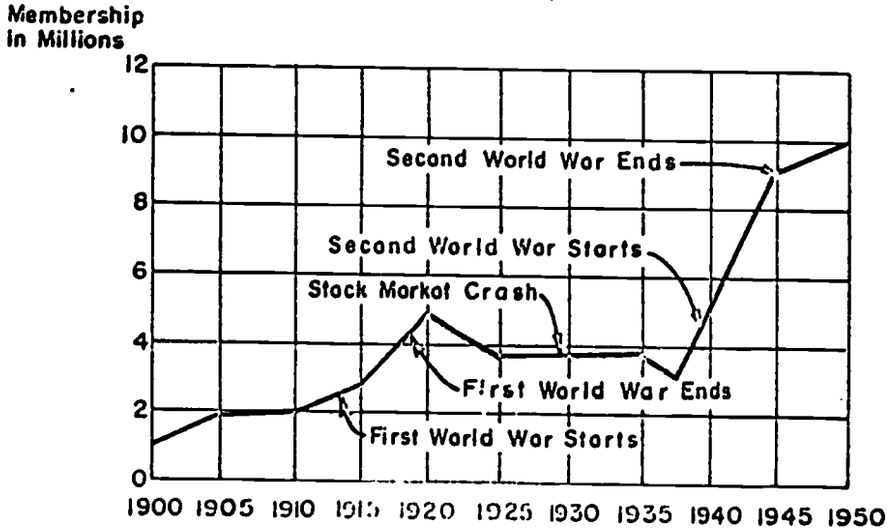
TRANSPORTATION ACCIDENT DEATH RATES 1965

Kind of Transportation	Death Rate Per 100,000,000 Passenger Miles
Automobiles and Taxis	2.40
Automobiles on Turnpikes	1.10
Buses	0.18
Railroad Passenger Trains	0.07
Scheduled Air Transport Planes (domestic)	0.38

23. According to the table above, which of the following was the safest kind of transportation in 1965 ?

(A) Automobiles on turnpikes
(B) Railroad passenger trains
(C) Buses
(D) Scheduled air transport planes

MEMBERSHIP OF AMERICAN FEDERATION OF LABOR 1900–1950



24. Which of the following statements about union membership is supported by the graph above?

(A) It increased most sharply during wartime.
(B) It increased most sharply just after a war.
(C) It decreased during the early years of a depression.
(D) It decreased just before a war.

Figure 2. Sample items in simplified form.

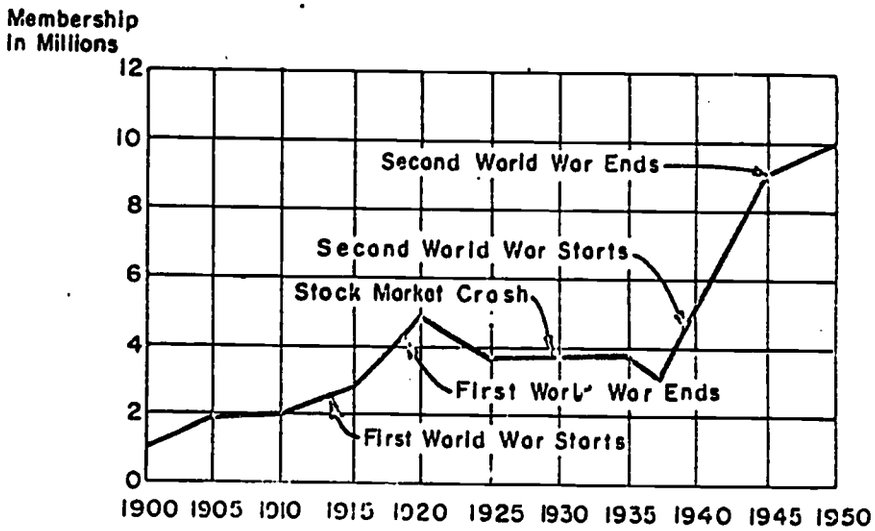
TRANSPORTATION ACCIDENT DEATH RATES 1965

Kind of Transportation	Death Rate Per 100,000,000 Passenger Miles
Automobiles and Taxis	2.40
Automobiles on Turnpikes	1.10
Buses	0.18
Railroad Passenger Trains	0.07
Scheduled Air Transport Planes (domestic)	0.38

23. What was the safest way to travel in 1965?

(A) Automobiles on turnpikes
(B) Railroad passenger trains
(C) Buses
(D) Scheduled air transport planes

MEMBERSHIP OF AMERICAN FEDERATION OF LABOR 1900-1950



24. The graph shows that the number of people in unions went

(A) up most during war
(B) up most just after a war
(C) down in the early years of a depression
(D) down just before a war

Figure 3. Mean scores on total test (48 items).

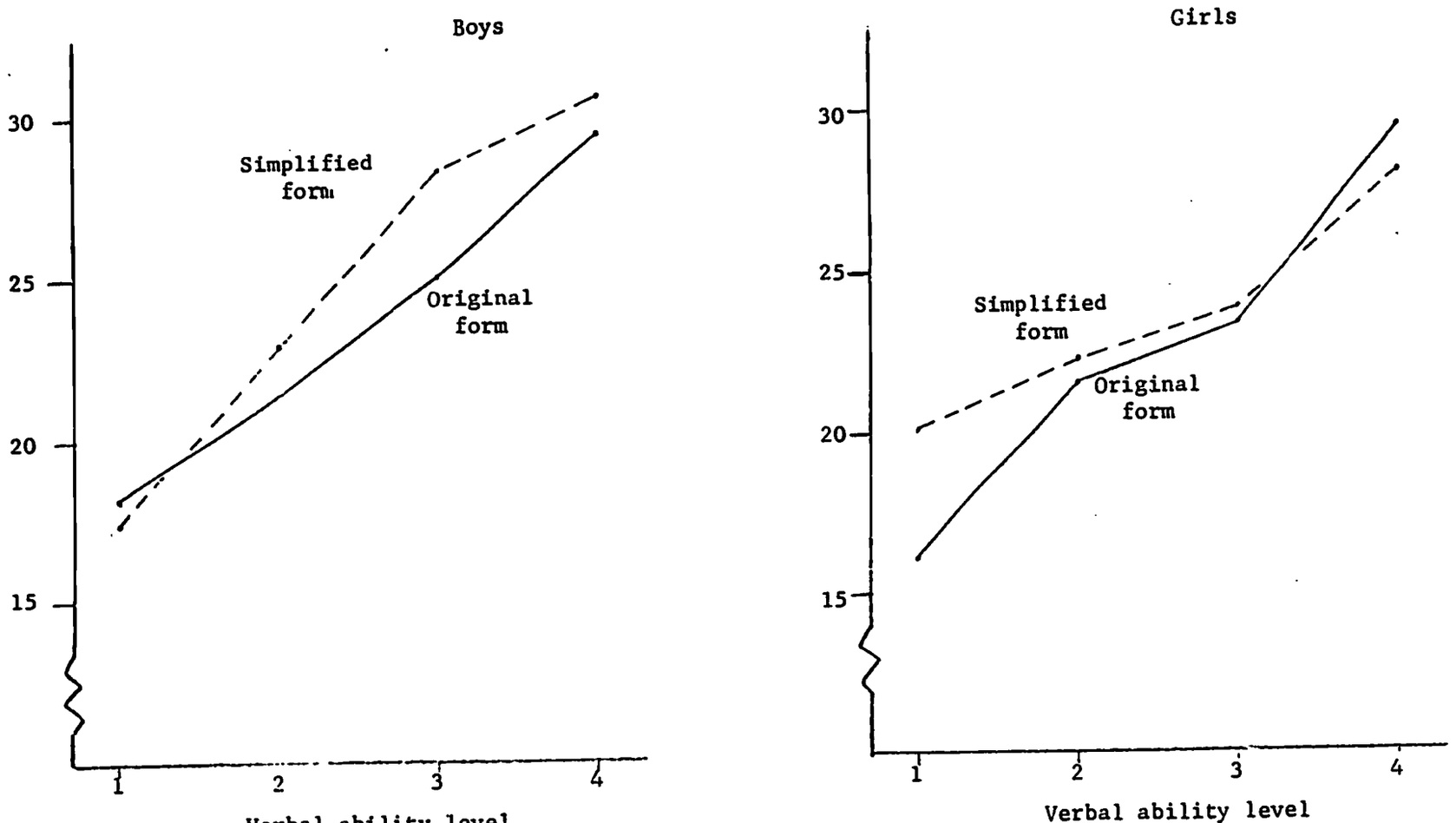


Figure 4. Scores on graphic and prose items.

Solid line = original items;
Broken line = simplified items.

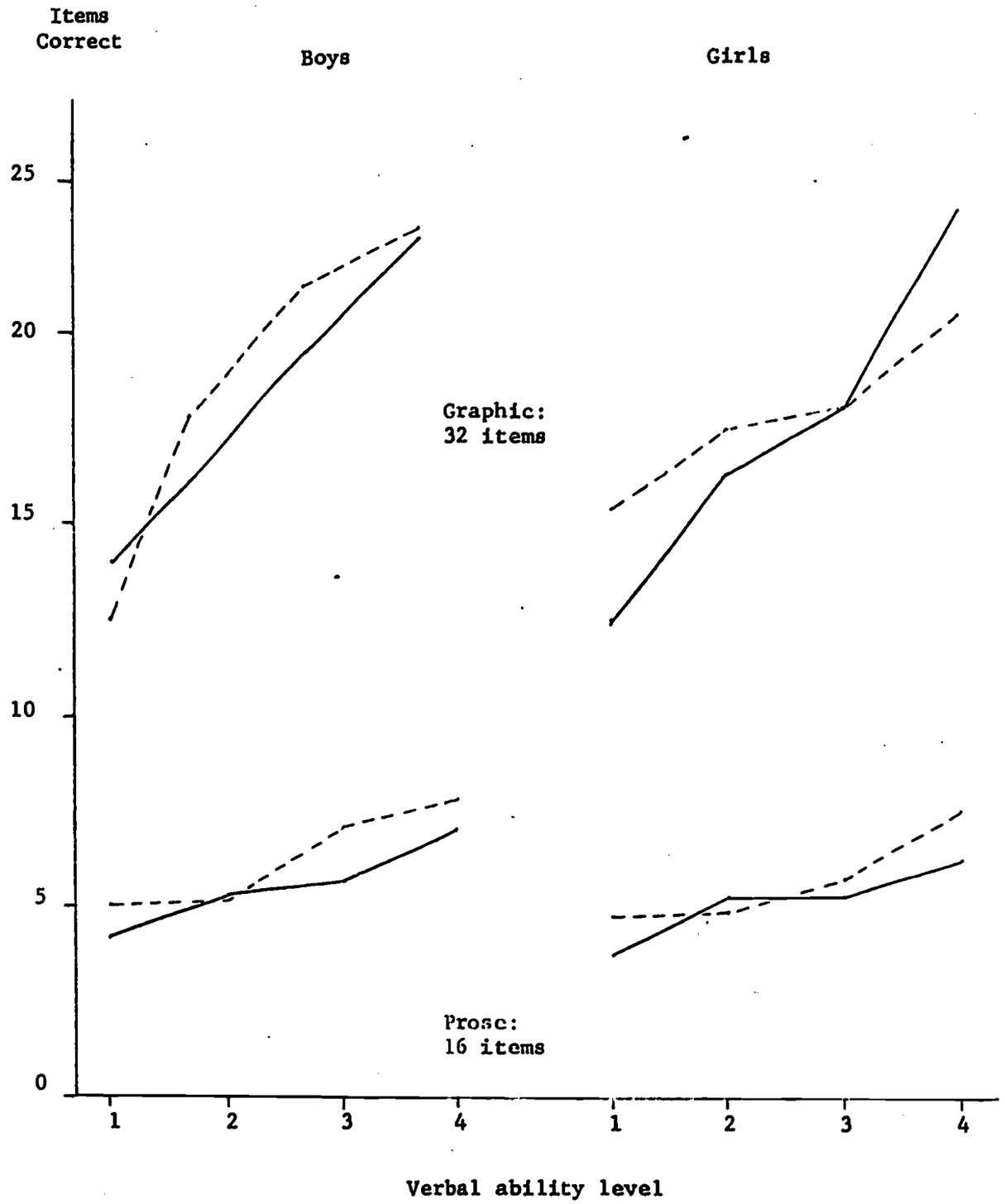


TABLE 1

Number of Students Participating

	Boys		Girls	
Test Form:	Original	Simplified	Original	Simplified
Verbal Ability:				
Level 1 (lowest)	34	29	17	19
Level 2	30	27	30	19
Level 3	24	23	20	21
Level 4 (highest)	28	24	21	23
Total	116	103	71	82

TABLE 2

Group Mean Scores on Full Test (48 Items)

	Boys		Girls	
Test Form:	Original	Simplified	Original	Simplified
Verbal Ability:				
Level 1	18.12	17.45	16.12	20.11
Level 2	21.40	22.96	21.54	22.26
Level 3	25.13	28.30	23.35	23.76
Level 4	29.46	30.58	29.48	28.00
Combined	23.16	24.38	23.10	23.76

TABLE 3

Group Mean Scores on Graphic Items (32 Items)

	Boys		Girls	
Test Form:	Original	Simplified	Original	Simplified
Verbal Ability:				
Level 1	13.97	12.45	12.41	15.42
Level 2	16.13	17.85	16.31	17.47
Level 3	19.50	21.26	18.10	18.10
Level 4	22.46	22.79	23.29	20.52
Combined	17.72	18.24	17.94	18.01

TABLE 4

Group Mean Scores on Prose Items (16 Items)

	Boys		Girls	
Test Form:	Original	Simplified	Original	Simplified
Verbal Ability:				
Level 1	4.15	5.00	3.71	4.68
Level 2	5.27	5.11	5.23	4.79
Level 3	5.63	7.04	5.25	5.67
Level 4	7.00	7.79	6.19	7.48
Combined	5.43	6.14	5.16	5.74

TABLE 5

F-ratios, Significance Levels, and Proportion of Variance Accounted For

Source of Variance	df*	Total Score	Graphic Items	Prose Items	Jr. High Graphic	Sr. High Graphic	Jr. High Prose	Jr. High Prose
Form (original vs. simplified)	1	$F = 3.24$	$F = 0.62$	$F = 6.03$ $p < .02$ $\eta^2 = .014$	$F = 0.09$	$F = 2.70$	$F = 8.01$ $p < .001$ $\eta^2 = .019$	$F = 1.34$
Ability	3	$F = 84.37$ $p < .001$ $\eta^2 = .401$	$F = 79.87$ $p < .001$ $\eta^2 = .385$	$F = 21.01$ $p < .001$ $\eta^2 = .145$	$F = 59.67$ $p < .001$ $\eta^2 = .321$	$F = 58.83$ $p < .001$ $\eta^2 = .314$	$F = 16.26$ $p < .001$ $\eta^2 = .117$	$F = 12.61$ $p < .001$ $\eta^2 = .092$
Sex	1	$F = 4.07$ $p < .05$ $\eta^2 = .006$	$F = 1.85$	$F = 3.88$ $p < .05$ $\eta^2 = .009$	$F = 3.61$	$F = 0.21$	$F = 0.10$	$F = 8.91$ $p < .01$ $\eta^2 = .022$
Form x Ability	3	$F = 0.58$	$F = 1.83$	$F = 1.26$	$F = 2.16$	$F = 1.82$	$F = 0.55$	$F = 2.29$
Form x Sex	1	$F = 0.90$	$F = 0.05$	$F = 0.62$	$F = 0.08$	$F = 0.01$	$F = 0.63$	$F = 0.17$
Ability x Sex	3	$F = 0.13$	$F = 0.09$	$F = 0.30$	$F = 0.15$	$F = 1.62$	$F = 0.87$	$F = 0.44$
Form x Ability x Sex	3	$F = 2.55$	$F = 3.86$ $p < .01$ $\eta^2 = .019$	$F = 0.37$	$F = 2.27$	$F = 5.76$ $p < .001$ $\eta^2 = .031$	$F = 0.38$	$F = 0.44$

*df within cells = 356

TABLE 6

Correlations of Estimated Reading Difficulty
With Actual Response Difficulty (N=48 Items)

	Original Items	Simplified Items	Change	Residuals
Entire Sample	.37	.16	.02	.06
Boys:				
1 (low)	.36	.26	.19	.25
2	.36	.19	.11	.16
3	.31	.05	-.30	-.34
4 (high)	.32	.02	-.16	-.21
Girls:				
1 (low)	.41	.16	-.05	-.03
2	.25	.11	.00	.11
3	.35	.24	.31	.27
4 (high)	.29	.11	-.07	-.05