AN EMPIRICAL COMPARISON OF THREE DISTRIBUTIONS OF ITEM DIFFICULTY WITH RESPECT TO THE RELIABILITY AND VALIDITY OF THE RESULTING MEASURES

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

Alfred J. Reynolds

1965

This is to certify that the

thesis entitled

AN EMPIRICAL COMPARISON OF THREE DISTRIBUTIONS OF ITEM DIFFICULTY WITH RESPECT TO THE RELIABILITY AND VALIDITY OF THE RESULTING MEASURES

presented by

Alfred J. Reynolds

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Ed. Psych.

Joe L. Saupe
Major professor

Date July 22, 1965

[redacted]

J ~~JUL 8 1970~~ /89

DOS

ABSTRACT

AN EMPIRICAL COMPARISON OF THREE DISTRIBUTIONS OF ITEM DIFFICULTY WITH RESPECT TO THE RELIABILITY AND VALIDITY OF THE RESULTING MEASURES

by Alfred J. Reynolds

The Problem

There is a discrepancy between the practice of test constructionists and that advocated by test theoriste. Most test theorists advocate that item difficulties be concentrated near the mean ability level of the examinees whenever it appears likely that item inter-correlations are low. However, in practice test constructionists continue to use items with a wide range of difficulty. It is the purpose of this study to determine which of three distributions of item difficulty, used in the construction of academic achievement tests, is most effective in terms of the homogeneity of test scores and their validity for grading purposes.

Procedure

Existing data from achievement tests were used to investigate the problem. Items from three term-end examinations were pooled. Items were selected from these pools to construct three 50 item experimental tests which represented a "Peaked", "Rectangular" and "Multimodal"

distribution of item difficulties for two subject areas. Reliabilities were computed and validities were determined, first by correlating total test scores with total instructor grades, and second by comparing the abilities of the tests to discriminate among criterion group means. Reliabilities and validities were compared statistically where possible and rationally where statistical comparisons did not seem appropriate.

Findings

1. The "Peaked Test" tended to have larger reliabilities than the "Rectangular Test" or the "Multimodal Test".

2. The "Rectangular Tests" tended to have larger reliabilities than the "Multimodal Tests".

3. When the validating criterion was total instructor grade, the "Peaked Tests" had larger validities than either the "Rectangular" or "Multimodal" tests.

4. The "Rectangular Tests" correlated higher with total instructor grades than the "Multimodal Tests".

5. The "Peaked Tests" discriminated most effectively between criterion groups when there was moderate interitem correlation.

6. When interitem correlations are high a greater spread of item difficulties will produce larger validities.

7. The "Rectangular Tests" were better discriminators than the "Multimodal Tests".

AN EMPIRICAL COMPARISON OF THREE DISTRIBUTIONS OF ITEM DIFFICULTY WITH RESPECT TO THE RELIABILITY AND VALIDITY OF THE RESULTING MEASURES

by

ohn
Alfred J. Reynolds

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

College of Education

1965

THI

ACKNOWLEDGEMENTS

The author wishes to acknowledge his indebtedness to Dr. Joseph L. Saupe, thesis director and committee chairman, without whose counsel, guidance, and assistance this investigation could never have been completed. Appreciation is also extended to Dr. Willard G. Warrington, guidance committee member, for his helpful suggestions and also to him and his staff at Evaluation Services, Michigan State University, who made possible the collection of data.

The investigator is also indebted to both Dr. John E. Hunter and to Dr. Edward B. Blackman for counsel, advice and willingness to serve on the guidance commitee.

A special word of gratitude is due to the author's wife, Bette, for her assistance, encouragement, and patience, and also to Nancy, Becky, Al and Amy, who still love their father.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

FIGURES Page

CHAPTER I

INTRODUCTION

There is a discrepancy between the practice of test constructionists and that advocated by test theorists. Cronbach and Warrington (1952), Gulliksen (1945), Richardson (1936) and Ebel (1959) are generally agreed that maximum precision of measurement will result from homogeneous item difficulties, concentrated near the mean ability level of the examinees. Myers (1962:565) says, "those who produce standardized tests continue to follow a tradition that items selected for a test should represent a wide range in difficulty." Noll (1957) states that " a test with adequate range of difficulty should include items ranging from quite easy to fairly difficult." The Office of Evaluation Services, at Michigan State University (1963) states that "it is generally desirable that item difficulty values vary from .20 to .80 so that an examination will discriminate at all levels." It is the purpose of this study to investigate this apparent paradox.

Review of Related Literature

The psychometric literature presents numerous articles dealing with the determination of appropriate distributions of item difficulty for specific purposes. Methods employed in investigation of this problem include rational analysis, empirical investigations with real data and empirical studies using hypothetical data.

Thurstone (1932) analyzed real data resulting from the construction and administration of numerous diagnostic spelling tests. The results indicated that tests composed of items concentrated at the 50% difficulty level would yield results most meaningful in the diagnoses of spelling difficulties.

Davis (1963:310) relying on the rational approach said that "we can see intuitively, however, that since many kinds of test items have low intercorrelations, a distribution of item difficulties clustered around the 50% level would often approximate the distribution required to obtain maximum discrimination throughout the range of scores."

Cronbach and Warrington (1952) analyzed hypothetical data in an attempt to determine the effect that spread of item difficulty would have on screening efficiency for various degrees of item reliability. Consideration was also given to the most appropriate level of item difficulty for maximum screening validity of a multiple choice test. The data consisted of conditional probability matrices mathematically manipulated to yield validity coefficients. The results indicated that the spread of item difficulty should vary directly as the intercorrelations of the items so that validity would be maximized over the whole range of scores. Validity at the extremes could be increased by either an increase in the spread of item difficulty or an increase in item unreliability.

Cronbach and Warrington (1952) assumed that most tests of mental ability have low intercorrelations between items and therefore recommended that "constructors of educational and psychological tests would be wise to make item difficulty constant in most of their tests, since this lowers validity only for persons having extremely high or low ability." (p. 147)

Ebel (1959) used a theoretical model and constructed hypothetical tests to analyze the relationship between item difficulty and examinee scores. He pointed out that it is not necessary, for effective measurement, to widen the spread of item difficulties as the range of ability of the examinees increases. Elsewhere (1954) he stated that "where item intercorrelations are low selection of items whose difficulty is near 50% tend to flatten the score distribution, to increase the dispersion of scores and thus to improve the discrimination power of the test as a whole."

Richardson (1936) suggested that if a certain percent of the examinees are to be sorted out, tests composed of items of the same difficulty will be more valid than those made of items which vary in difficulty. He also indicated that these items should have a difficulty level which corresponds to the ability level of the examinees in the group that are to be accepted. Cronbach and Warrington (1952) also supported this position when they stated that "in order to design a test which rejects the poorest F percent of the men tested, items should on the average be located at or above the threshold for men whose true ability is at the Fth percentile." (p. 147)

The literature, referred to above, reveals that test theorists have been concerned chiefly with two types of problems. One was to build tests for selection purposes, which would divide a group into two parts, i.e. those to be accepted versus those to be rejected. The other problem was to discriminate among the examinees over the entire continuum of the specified behavioral characteristic.

A third problem is reflected in the following literature. It is concerned with the need for achievement tests which will separate a group of examinees into subgroups for marking purposes. Davis (1950:311) says that "the assignment of marks (which calls for the division of a group into several parts) demands maximum accuracy of measurement at the several dividing points scattered along the range of scores."

In regard to the construction of tests for various purposes, Gulliksen (1945:91) states that "whether it is actually best to concentrate all items at one difficulty level, --- or to distribute items over a difficulty range in accordance with present test practice, can be determined only by experiments such as those reported by Thurstone (1932) and Richardson (1963)."

Jackson (1952) engaged in such a study in an effort to develop a practical method for selecting items for a test which would separate examinees into sub groups for marking purposes so that there would be a minimal error of measurement at the critical division points. He developed an item analysis procedure which employed the use of chi-square for selecting items which discriminated between adjacent groups. This procedure was validated with the use of data from achievement tests given at Michigan State University. The results indicated that the "adjacent group technique" proved a satisfactory method for selecting test items under the conditions of this experiment.

Two limitations are apparent in this study. First, there was no real external criterion to use in computing a validity coefficient. Second, the ratio of items selected to those needed was too low (less than 7%) for this technique to be seriously considered as a procedure for practical test construction.

Myers (1962) attempted to ascertain which of two types of item difficulty distributions would produce greater reliability and validity. He used the items from a 150-item test designed to predict the scholastic aptitude of college students to construct four tests of 24 items each. Two "Peaked Tests" consisted of items whose difficulty was between 40% passing and 70% passing. The other two tests were "U shaped" and consisted of items whose difficulty was outside the range of those on the "Peaked Test". Samples from twelve colleges were selected. Reliabilities were computed for each sample by correlating the two 24 item tests of each type. Validities resulted, in each sample, from correlating the test scores of freshman with their average college grades. The results failed to show any statistically significant difference between the validities for the two types of distributions. The hypothesis that "Peaked Tests" yield the best reliability was given tentative support.

Several factors may have contributed to the inconclusiveness of Myers' (1962) study. The criterion, used to compute validities, consisted of the average grades received by freshman. These students were selected from twelve different liberal arts colleges which differed widely, both academically and geographically. The sample size from each institution varied from 37 to 392. The samples were also selected in different ways. Reliabilities and validities were computed for the test results for each sample and were then compared for the two tests by using Wilcoxon's matched pairs signed test.

Meyers considered each sample to be a matched pair since the same individuals in each sample responded to both tests. The assignment of equal weights, in the significance tests, to the results of the samples which differed so much in size, appears to be questionable.

Other limitations to the study are the short, 24 item, tests used to compute reliability, and finally that the items were chosen solely on the basis of difficulty indices with only a reference, which is not clearly defined in the report, to a discrimination index.

The College Achievement Test

The college objective achievement test is becoming increasingly popular as a means for determining grades for students. The Committee on Measurement and Evaluation of the American Council of Education emphasizes this importance. It states that "test data frequently are not the sole determiners of course grades, but normally they make a relatively major contribution. --- the final examination may carry equal or greater weight than other work." (1959)

The objective standardized test has grown in popularity among college personnel to the extent that Educational Testing Service is now in the process of developing "Course Examinations, intended to measure end-of-course achievement in widely taught undergraduate courses, including technical and professional subjects offered for college credit." (1963) Innovations such as educational television, independent study, programmed teaching, team teaching, etc., have forced attention to the use of achievement examinations as evaluative devices.

Larger universities often use the scores from objective achievement tests to assist in assigning final course grades, or to give credit in lieu of taking specified basic courses within the university. Michigan State University is one which uses objective achievement test scores to determine 50% of the letter grade for thousands of students in Natural Science, Social Science, Humanities, and American Thought and Language courses.

Theoretical Implications of Previous Research

The psychometric literature indicates that appropriate item difficulty distribution is a function of the use that is to be made of the test scores. Three types of item difficulty distribution are indicated for using achievement tests in the assignment of grades. First, a "Reetangular" distribution results from the common sense approach. It is assumed that there is a rather wide range of ability among examinees and that these various levels of ability require appropriate levels of item difficulty for proper discrimination. Items must range from the very easy to the very difficult and include every level of ability. The frequency distribution of these items difficulties appears rectangular in shape since there are few items in each category but there are many categories corresponding to the many levels of ability.

The "Peaked" distribution has received most theoretical and experimental support. It is assumed that if item intercorrelations are low, an item of 50% difficulty will make more discriminations than an item of any other difficulty. By concentrating all of the test items as near this level of difficulty as possible, it is assumed that maximum discrimination will result over the whole range of ability levels.[1]

If the test is to be used to divide a group into well defined sub groups, discrimination among individuals near the critical division points is imperative (Davis 1963; Jackson 1952). Test theory implies that item difficulty be concentrated at these points, according to

[1] A "normal" distribution of item difficulties would be intermediate between rectangular and peaked. Hence, by interpolation the results of this study may be tentatively extended to other test types.

Richardson (1936) and Davis (1963). A threefold problem is thus presented to the test constructor who wishes to build a test which will function in this manner. First, he must determine the position of the critical division points on the continuum of test scores. Second, he must determine the appropriate number of items to concentrate at each level. Finally, the difficulty levels of these items must be determined.

A solution to the first problem can be found in the assumption that a certain percentage of the examinees will receive a particular letter grade. If the desired percentages of students receiving each grade can be determined, then these percentages indicate the appropriate point of division on the score continuum.

The second problem is solved by the following reasoning. It is commonly accepted that the reliability of a test is a function of test length, provided that all items are somewhat equally effective. Discrimination among a larger number of scores concentrated about a given score should require greater precision of measurement than would be the case where discrimination is necessary among a smaller number of scores concentrated in a given segment of the score continuum. It follows then, that the number of items concentrated at each level must be proportional to the number of scores expected near this level. The proportion of each letter grade indicates tne proportion of students to be retained in each score category. Knowing the number of items to be included in a particular test, the appropriate number of items to be selected for each level can be computed from equation (1).

$$G \; X \; I = N \tag{1}$$

where:

G = the percent of students receiving each letter grade

I = the total number of items included on the test

N = the number of items desired for a particular grade level.

A solution to the problem of the appropriate item difficulty to concentrate at each division point is suggested by Lord (1952), Cronbach and Warrington (1952) and Davis (1963). The consensus is that if a given percent of a group is to be selected item difficulty should be near that corresponding to the percentage of examinees to be retained. The percentage of students receiving a particular letter grade indicates the percentage of students to be retained in each category. But, students are also to be retained for all categories above the one in question. Therefore, the total percentage of students retained at each division point can only be determined by summing the percentages of students in all categories above that point. This percentage will also indicate the difficulty of the items to be concentrated at that particular point, assuming that items of a given degree of difficulty discriminate most effectively at a corresponding degree of ability.

The commonly accepted phenomenon known as "regression toward the mean" implies that true item difficulties for a group of examinees will be located in the direction of the mean from the actual item difficulties computed from a sample group (Hayes, 1963). Most effective discrimination for groups should result, therefore, from a limited range of item difficulty focused about the various division points, but skewed toward the mean.

A theory has been developed here which incorporates principles of concentration and spread of item difficulty. Groups of items are concentrated but the groups are of different size and are spread out in order to discriminate more adequately at different levels of ability. The application of this theory will result in the construction of a "Multimodal" test.

The Setting

The University College at Michigan State University is designed to provide for each student a common core of courses in general education. These courses include those fundamental areas of knowledge which are felt to be an important part of the education of all students regardless of the individual's field of specialization. All undergraduates are, therefore, required to take a sequence of courses in American Thought and Language, Natural Science, Social Science and Humanities.

All students enrolled in these University courses are required to take a term-end examination which constitutes 50% of the final grade received in the course. These examinations are standardized achievement tests prepared from items submitted by instructors in the various courses and assembled under the direction of the Office of Evaluation Services. All students enrolled in a particular course for a given quarter are tested with the same instrument. Number grades are assigned solely on the basis of the scores received on these tests. These number grades are then averaged with a number grade assigned by the instructor to determine the final letter grade assigned in a particular course.

The instructor grades are assigned completely independently of the test scores. They are based on the student's performance with regard to his instructor's assignments, tests, recitations, etc. The instructor number grade is assigned on a 15 point scale. It may be converted into a letter grade by use of the following code:

1, 2, or 3 equal F; 4, 5, or 6 equal D; 7, 8, or 9 equal C; 10, 11, or 12 equal B; 13, 14, or 15 equal A.

The Problem

Since the objective achievement test plays an essential role in the determination of college student's grades, it is important that these tests be constructed in such a way as to make their results function most efficiently. It has been shown that a significant factor in determining the efficiency of a test, for a specified purpose, was the distribution of item difficulty.

It is, therefore, the purpose of this study to compare the effectiveness of using three different distributions of item difficulty, in the construction of academic achievement tests, in terms of the homogeneity of the scores and their validity for grading purposes.

This study used achievement test data, available from University College term-end examinations at Michigan State University, to investigate the problem. Three experimental tests were constructed for each of two subject areas. These tests represented three different types of item difficulty distributions, namely, (1) "Peaked", (2) "Rectangular", and (3) "Multimodal". The relative effectiveness of these tests was judged by:

1. The level of internal consistancy as determined by Kuder Richardson #20,

2. The degree of correlation with instructor grades, and

3. The ability to discriminate among instructor-grade groups.

The item difficulties used in this study were based on a stratified sample of fifty students in the upper twenty seven percent of the distribution of total test scores and a stratified sample of fifty students in the lower twenty seven percent of the distribution of total test scores. The samples of fifty students were chosen so that they possessed approximately the same distributions of scores as the larger groups from which they were chosen. Item difficulties consisted of the total proportion of students answering the item correctly in both the upper and lower sample groups.

Prospectus

The following chapter outlines the general plan of the experiment. The initial test data and subjects are discussed and the "Peaked", "Rectangular" and "Multimodal" experimental tests are described. Chapter III presents and discusses the analyses of the results of the experimental tests. Reliabilities are compared rationally and validities are statistically compared both as to correlation with instructor grades and as to ability to discriminate among grade groups. The final Chapter summarizes the procedure and findings of the investigation. It also points out the limitations of the study and offers conclusions and recommendations.

CHAPTER II

PROCEDURES

The purpose of this chapter is to outline the general plan of the experiment. The initial test data and subjects used in the study will be described. Finally, the experimental tests developed for, and used in, the study will be discussed.

General Procedure

Available data from achievement tests were used to investigate the relative effectiveness of the three item difficulty distributions in separating examinees into groups for grading purposes. Term-end examinations from two subject areas were selected. A rather large pool of items was needed in order to build an experimental test of the required item difficulty distribution and also of a satisfactory length. The items from three term-end tests given in sequence in each subject area, to the same students were, therefore, combined to form item pools. Items for the experimental tests were taken from these pools.

The tests selected were those which had been given in Natural Science and in Social Science for three successive terms, i.e., Fall 1963, Winter 1964 and Spring 1964. Students normally take the three courses in sequence, starting in the Fall and finishing in the Spring.

Item analyses were obtained for each of these tests. Difficulty and discrimination indices were taken from these analyses. The difficulty indices were used in selecting items for the experimental tests. The discrimination indices were used to reject undesirable items and to keep the items of each test as similar as possible in regard to this stastistic.

The item discrimination index available on the items used in this study, was determined by use of the table prepared for this purpose by Flanagan (1936). This index is an estimate of the product moment correlation coefficient between an item and the total test score. The proportion of successes in both the lowest and highest 27 percents of stratified random samples of examinees were used in entering Flanagan's Table.

The students used in the study were those who had taken all three terms of each subject in the proper sequence; Fall 1963, Winter 1964 and Spring 1964, and had also received an instructor grade for each term. Answer sheets for these students were obtained and rescored for each of the three experimental tests. The scores from each type of experimental test for all three terms were combined to yield a total score.

Reliability coefficients were computed for each of the experimental tests by using the Kuder Richardson #20 formula. Rational comparisons were made between these reliability coefficients. They were also corrected for length and compared with the reliabilities of the original tests.

Instructor grades for each student for each of the three terms were obtained and added together for a total instructor grade. These composite instructor grades served as the criterion for comparing the validities of the experimental tests. Validities of a first type were estimated by the product moment correlation coefficient computed between total instructor grade and total test score. Statistical tests were used to compare these validities.

As a second validity analysis, groups of students who had received an average instructor grade of A, B, C, or D, for all three terms were identified. Test score means were computed for each of these groups for each of the three experimental tests. Adjacent group means were compared statistically in order to determine which experimental test most adequately discriminated among instructor grade groups.

Selection of Tests

The term-end examinations used in this study had been given to large numbers of students in three consecutive courses in the Natural Science and Social Science sequences. The series consisted of the term-end examinations for Fall 1963, Winter 1964 and Spring 1964 for both courses.

The Natural Science examinations each contained 125 items for a total pool of 375 items. The Social Science tests contained 100 items in the Fall, 110 in the Winter, and 120 for the Spring quarter. This gave a total of 330 items for the Social Science item pool. The average reliability as determined by Kuder Richardson #20, was .87 for the Natural Science tests and .80 for the Social Science tests. The average validity, which represented a correlation with instructor grades, was .71 for the Natural Science tests and .62 for the Social Science tests.

After elimination of items in the Natural Science item pool which had discrimination indices of less than .25; 250 items remained. The difficulty indices of these items ranged from .09 to .97. The frequency distribution of the item difficulties for these items is given in Table I.

Social Science items were eliminated from the item pool if they had a discrimination index less than .20. This reduced the pool to 251 items. The item difficulties of these items ranged from .08 to .98. The frequency distribution of these items difficulties is also given in Table I.

TABLE I. Frequency Distributions of Item Difficulties for Natural Science and Social Science Item Pools.

Difficulty	Natural Science	Social Science
.86-.98	19	35
.76-.85	45	43
.66-.75	52	47
.56-.65	58	45
.46-.55	43	38
.36-.45	31	31
.26-.35	16	14
.16-.25	13	3
0-.15	4	3
TOTAL	250	251

Both the Natural Science and Social Science examinations were essentially power tests. Even though a time limit was imposed, most of the examinees responded to every item. Both tests were of the multiple choice type. The Natural Science test had five choices for each item. The Social Science items each had four choices. Answers were recorded on IBM form I.T.S., 1000 B 4701. Test papers were carefully checked for marking more than one answer per item. These were excluded from the study. The remaining answer sheets were then scored on an IBM scoring machine. The score on a test was the total number of correct responses since no correction was made for guessing.

Selection of Students

Students who were enrolled in the Natural Science sequence; N.S. 181 - Fall 1963, N.S. 182 - Winter 1964, and N.S. 183 - Spring 1964, at Michigan State University comprised the subjects for part of this study. The remaining subjects were those students who enrolled in the Social Science sequence; S.S. 231 - Fall 1963, S.S. 232 - Winter 1964, and S.S. 233 - Spring 1964.[1] Students who did not take all three examinations, who did not receive an instructor grade for all of the courses in the sequence, or who used Form B on the Fall term-end examination in both subject areas, were eliminated from the study. Most of the students were college freshman and sophomores.

There were 5168 students who took the Fall Natural Science examination, 4408 took the Winter test, and 3371 were tested at the end of the Spring quarter. A total of 1423 students were available who had taken all three Natural Science examinations, received an instructor grade for each course and used Form A on the Fall term examination.

[1] Some students may have been in both courses.

The term-end examinations in Social Science were taken by 3189 students in the Fall, by 2698 students in the Winter, and by 2295 students in the Spring. Of these, 909 students were available, who had taken all three examinations received instructor grades for each course, and used Form A in the Fall term examination.

The Experimental Tests

Three experimental tests of 50 items each were developed for use in both the Natural Science and the Social Science areas. The items for these tests were taken from the respective item pools which resulted from combining the items of the term-end examinations in these two subjects. Items were chosen which had the largest discrimination index and the appropriate difficulty level for the test being developed. Item discrimination indices were balanced as much as possible. Attention was also given to balancing the number of items taken from the Fall, Winter, and Spring term examinations. Tables II and III present the resulting distribution of item difficulties for the three tests in each area.

Table IV shows the mean item difficulties, item discrimination indices, and indicates the number of items taken from the Fall, Winter, and Spring term-end examinations.

TABLE II. Frequency Distributions of Item Difficulties for Experimental Tests in Natural Science.

Difficulty Range	Peaked Test	Rectangular Test	Multimodal Test
.86-1.00	-	7	9
.76- .85	-	6	20
.66- .75	-	6	-
.56- .65	25	7	-
.46- .55	25	7	-
.36- .45	-	6	15
.26- .35	-	6	-
.16- .25	-	6	2
0- .15	-	-	4
TOTAL	50	50	50

The Peaked Test

It was possible to construct "Peaked" tests for both subject areas from the item pools. The items ranged in difficulty from .46 to .63 for the Natural Science test and from .48 to .63 for the Social Science test. The composition of these tests with respect to item difficulty, discrimination index and the designation of the test from which the item came, are given in the Appendix.

TABLE III. Frequency Distributions of Item Difficulties for Experimental Tests in Social Science.

Difficulty Range	Peaked Test	Rectangular Test	Multimodal Test
.86-1.00	-	6	10
.76- .85	-	8	22
.66- .75	-	5	-
.56- .65	25	7	-
.46- .55	25	9	-
.36- .45	-	7	10
.26- .35	-	6	3
.16- .25	-	2	2
0- .15	-	-	3
TOTAL	50	50	50

The Rectangular Test

Sufficient items were available in the item pools to construct a "Rectangular Test" for each subject area. Few items having the same difficulty index were used in either test. The range of difficulty for the Natural Science test was from .23 to .93. For the Social Science test it was from .22 to .92. The composition of the tests with respect to item difficulty, discrimination index, and source of items is given in the Appendix.

TABLE IV. Mean Difficulties, Mean Discrimination Indices and Source of Items for Experimental Tests.

Test	Mean Difficulty	Mean Discrim.	Fall	Winter	Spring
NATURAL SCIENCE					
Peaked	.53	.48	17	16	17
Rectangular	.56	.46	16	16	18
Multimodal	.64	.44	16	16	18
SOCIAL SCIENCE					
Peaked	.55	.39	16	16	18
Rectangular	.58	.40	17	15	18
Multimodal	.64	.37	16	14	20

The Multimodal Test

Consistent with the theory for constructing a "Multimodal Test" (p. 9-12), a test of this nature was constructed for each subject area. The instructor grades used in the construction of these tests were assigned independently of the term-end examinations in the basic college courses. They were reported on a 15 point scale. A small percentage of the grades other than these, i.e., deferred or incomplete were excluded from the study. These percentages were averaged for the three quarters involved in the study and these data are given in Table V and VI.

TABLE V. Percentage of Students Receiving Each Instructor Grade for Natural Science.

Grade	Fall	Winter	Spring	Average
A	12.0	12.2	11.0	11.7
B	28.3	29.9	28.8	29.0
C	39.6	41.6	41.3	40.8
D	14.2	12.9	15.1	14.1
F	4.1	2.7	2.9	3.2

TABLE VI. Percentage of Students Receiving Each Instructor Grade for Social Science.

Grade	Fall	Winter	Spring	Average
A	9.2	9.1	8.8	9.0
B	24.8	24.5	26.9	25.4
C	44.6	45.2	43.5	44.4
D	15.6	16.0	16.5	16.3
F	4.9	4.5	3.4	4.3

Tables VII and VIII present the number of items and the respective difficulty level needed to construct a 50-item "Multimodal Test" for Natural and Social Science. Average percentage of instructor grades were taken from Tables V and VI. The number of items needed in each category was computed from equation (1), (p. 11)

TABLE VII. Percentages of Students Receiving Instructor Grades and Number of Items Needed at Each Level of Difficulty for the Natural Science Multimodal Test.

Instructor Grade	Percentage Receiving	Number of Items	Difficulty Level
A	11.7	6	12
B	29.0	15	41
C	40.8	20	82
D	14.1	9	96

TABLE VIII. Percentages of Students Receiving Instructor Grades and Number of Items Needed at Each Level of Difficulty for the Social Science Multimodal Test.

Instructor Grade	Percentage Receiving	Number of Items	Difficulty Level
A	9.0	5	9
B	25.4	13	34
C	44.4	22	79
D	16.3	10	95

In the actual construction of the test the distribution of item difficulties was skewed toward the mean. The resulting ranges of item difficulty for each indicated level are given in Table IX. The data concerning the actual items included in both tests are given in the Appendix.

TABLE IX. Range of Item Difficulty in Each Category for the Multimodal Tests.

NATURAL SCIENCE		SOCIAL SCIENCE	
Indicated Difficulty	Actual Range of Difficulty	Indicated Difficulty	Actual Range of Difficulty
.96	.89-.97	.95	.91-.95
.82	.78-.82	.79	.75-.79
.41	.41-.45	.34	.35-.40
.12	.12-.20	.09	.08-.22

Summary

In this phase of the investigation the plan of the experiment was considered. The term-end examinations used in the study were discussed. The development of, and the characteristics of the three experimental tests, "Peaked", "Rectangular", and "Multimodal" were described.

CHAPTER III

RESULTS

The purpose of this chapter is to present and discuss the results of the analyses described earlier, from using the three experimental achievement tests in two different subject areas. Each of these tests represents a different type of item difficulty distribution, namely, (1) "Peaked", (2) "Rectangular", and (3) "Multimodal". Each test was constructed and scored from data available on term-end achievement examinations at Michigan State University.

Reliabilities of Experimental Tests

Reliabilities for the experimental tests were computed from the formula developed by Kuder and Richardson and reported by Gulliksen (1962). This formula is:

$$r_{xx} = \left[\frac{K}{K - 1}\right]\left[1 - \frac{\sum_{g = 1}^{K} s_g^2}{s_x^2}\right]$$

where r_{xx} is the reliability coefficient of the test,

K is the number of items in the test,

s_g^2 is the variance of item g (equals p_g $(1-p_g)$ where p is the percentage getting the item correct), and

s_x^2 is the test variance.

The item difficulties used in constructing the experimental tests were considered to be good estimators of the percentages of examinees getting each item correct. These item difficulties can be found in the Appendix and were used in the computation of the reliability coefficients.

The use of these item difficulties in the Kuder Richardson #20 formula seems justified if it can be assumed that the method used to compute them is defensible, that the average ability levels of the three groups used in these computations are approximately equal to that of the experimental group, and that variance of item difficulties is not seriously affected.

Flanagan (1939) defended the method used in computing the item difficulties (see page 15) for he said that, "In practice it appears that frequently it is satisfactory to use the values obtained from this chart together with an index of difficulty found by averaging the difficulties for the upper and lower groups".

Although a number of students with low ability failed to complete the sequence of courses used in this investigation, a number of those with superior ability also elected not to complete the entire sequence, since they passed examinations in lieu of taking the final courses in the sequence. Attrition at both ends of the ability continuum is also indicated in Tables V and VI. It can be seen that there is a decline in the percentage of both A's and F's received, from Fall term to Spring term in both subject areas. This attrition of the top and low ability groups did not greatly affect the average ability of the groups and, therefore, for the purpose of this study, the average abilities of the three groups were considered to be the same.

It is evident from the Kuder Richardson #20 formula that test reliability is dependent upon the variance of item difficulties and Tucker (1949) has indicated that while the mean item difficulty might change the variance probably would not. He said that "Estimates can be made of item variance from experimental forms or that it might even be possible to guess a practical value of item variance from editorial judgement."

Since the assumptions regarding method of computation of the original difficulty indices, equality of the groups involved and variance of item difficulties did not appear to be seriously violated, item difficulties used in constructing the experimental tests were used in the Kuder Richardson #20 reliability formula. The resulting reliability coefficient appear in Table X. The test means, and standard deviations for the experimental tests are also given in Table X.

TABLE X. Means, Standard Deviations, and Reliabilities of the Experimental Tests.

Test	Mean	Standard Deviation	Reliability
		NATURAL SCIENCE	
Peaked	27.74	7.53	.80
Rectangular	26.89	5.41	.68
Multimodal	31.77	4.64	.63
		SOCIAL SCIENCE	
Peaked	27.97	6.14	.69
Rectangular	28.78	5.69	.70
Multimodal	32.39	4.41	.59

Evidence that the item difficulties operated as expected can be ascertained by an inspection of the experimental test score means as they appear in Table X, and comparing them with the average difficulties of these tests as they appear in Table IV. (page 25) The means of the tests tend to descent in order of magnitude from a high in the "Multimodal Test" to a low in the "Peaked Test". This tendency is a reflection of the fact that the average item difficulty for the tests vary in the same direction. The "Multimodal Test" was easiest with a mean item difficulty of .64 while the other tests were more difficult, having mean item difficulties in the low .50's. Although there is a reversal in this tendency in the Natural Science area involving the "Peaked" and the "Rectangular" tests, this reversal is undoubtedly only apparent since there is no significant difference, at the .05 level for a two-tailed test, between the means of the scores for these two tests. (t = .014; d.f. = 1422)

The reliability coefficients appearing in Table X are indices of internal consistancy, computed on the same sample, and the author is aware of no statistical procedure for determining whether or not the differences among them 'are statistically significant. A rational analysis of data pertaining to the reliabilities of the experimental tests will, therefore, be presented.

Inspection of Table X reveals a general tendency for the reliabilities to descend in order of magnitude from the "Peaked Test" to the "Multimodal Test" with the reliability of the "Rectangular Test" falling between these two. The pattern of the standard deviations of the experimental tests supports this observation, since they descend consistantly, for both subject areas, in the same order suggested by the reliabilities. This is even true in the Social Science area where the standard deviation of the "Peaked Test" exceeds that of the "Rectangular Test" even though the magnitude of the reliability coefficients is reversed thus reflecting the fact that total item variance for the "Peaked Test" was also greater. The rank order of the size of the standard deviations of the experimental tests supports the hypothesis that the "Peaked Tests" were most reliable since it is generally true that a test which spreads out examinees farthest on the score continuum, is most precise in measuring the amount of the behavioral characteristic being assessed. (Saupe 1961)

There is one discrepancy in the general pattern of the reliability coefficients. In the Social Science area the "Rectangular Test" has a larger reliability coefficient than the "Peaked Test". This difference is small however, being only .01. The actual difference in favor of the "Peaked Test" in the Natural Science area is twelve times as large as the difference in favor of the "Rectangular Test" in the Social Science area.

An analysis of the function that discrimination indices have in determining test reliability is also relevant to the interpretation of the discrepancy in the Social Science Area. Gulliksen (1962:379) has shown that "the reliability of the test can be increased only by making the average item variance smaller or the average item reliability index larger", and has presented the following formula showing the relationship:

$$r_{xx} = \left[\frac{K}{K - 1}\right] \left[1 - \frac{\overline{(s_g^2)}}{K(\overline{r_{sg}s_g})^2}\right]$$

where K is the number of test items,

$\overline{(s_g^2)}$ is the average item variance, and

$\overline{r_{xg}s_g}$ is the average item reliability index.

Table IV (p. 25) reveals that the average discrimination index for the "Rectangular Test" in the Social Science areas is .01 larger than that of the "Peaked Test". It seems reasonable to conclude that this increase in the average discrimination index would result in an increase in the average reliability index.

According to Gulliksen, this increase in the average reliability index would function to increase the reliability of the "Rectangular Test" over that of the "Peaked Test". The average item variance was also smaller for the "Rectangular Test" thus it too functioned to increase the reliability of this test. These two variables both operated in the same direction and produced an increase of only .01 in the reliability of the "Rectangular Test" in the Social Science area. The meager influence of the small differences in average discrimination index and average item variances were reflections of the fact that average item variance was free to vary only from near 0 to .25, average item standard deviation from near 0 to .5, and average item discrimination index from .20 to .68 making it possible to have average item reliability indices somewhere between .10 and .34. It was, therefore, concluded that small changes in parameters could have but little influence on total test reliability as long as the number of items remained constant. It could also be concluded that although the average discrimination index of the "Peaked Test", in the Natural Science area, was .04 higher than the "Rectangular Test", this would not function to account for the .12 difference in the reliabilities of these tests as shown in Table X.

Comparison of the reliabilities of the experimental tests with the average reliabilities of the original tests presents difficulties beyond the lack of the statistical test indicated earlier. Items with low indices of discrimination were systematically eliminated from the experimental tests and this fact alone should have caused them to have higher reliability coefficients. Statistical comparison is also hampered by the differences in length between the experimental tests and the originals. In spite of these limitations, a rational comparison between them is indicated in order that the original tests might serve as bench marks for evaluation of the experimentaltests.

In order that this comparison be made as meaningful as possible, the reliabilities of the experimental tests were adjusted for length by using the Spearman-Brown formula developed for this purpose and reported by Cronbach (1960). This formula is as follows:

$$r_n = \frac{nr}{1 + (n - 1)\ r}$$

where r_n is the reliability of the lengthened test,

r is the reliability of the original test, and

n is the ratio of the new test length to that of the original test.

For the Natural Science experimental tests, "n" became 2.5 since the original tests each contained 125 items while the experimental tests consisted of 50 items. "n" was set equal to 2.2 for the Social Science experimental tests, since the average length of the original tests was 110 items, while the experimental tests contained 50 items each. Reliabilities resulting from these computations are given in Table XI along with the average reliabilities of the original tests.

The statistics given in Table XI indicate that only the reliability of the "Peaked Test" exceeded that of the original test in the Natural Science area. In

In the Social Science area both the "Peaked Test" and the "Rectangular Test" had reliabilities of greater magnitude than the original tests.

TABLE XI. Reliability Coefficients for Original and Experimental Tests Adjusted for Length.

	Original Test	Peaked Test	Rectangular Test	Multimodal Test
NATURAL SCIENCE	.87	.91	.84	.81
SOCIAL SCIENCE	.80	.83	.84	.76

Validities of the Experimental Tests

One of the criteria for judging which of the three experimental tests discriminates most effectively, is the correlation with instructor grades. Pearson product-moment correlation coefficients were computed between total instructor number grade and total test score for each of the three experimental tests. (see p. 18) These coefficients are given in Table XII along with the differences between them.

TABLE XII. Validities, Differences Between Them and t's for these differences, for the Experimental Tests in Natural Science and Social Science.

	P	R	M	(P - R)	t	(P - M)	t	(R - M)	t
N.S.	.81	.75	.61	.06	6.46	.20	20.42	.14	12.38
S.S.	.65	.59	.49	.06	3.26	.16	8.29	.10	4.67

A statistical test for the significance of differences between correlation coefficients, has been reported by Lindquist (1940). This test is appropriate where two or more tests have been correlated for the same group of subjects with the same variable. In the present study the three experimental tests were all correlated with the same student's instructor grades. This test of significance was, therefore, applied to the differences between the validity coefficients of the experimental tests used in this investigation.

The "t's" given in Table XII were used to test the null hypotheses that the population correlations between instructor grades and test scores were the same for each pair of experimental tests. Following is the formula used.

$$t = \frac{(r_{12} - r_{13}) \sqrt{n - 3} \sqrt{1 + r_{23}}}{\sqrt{2} \sqrt{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2(r_{12})(r_{13})(r_{23})}}$$

where r is the correlation between two variables, and

n is the number of cases.

Data used in the computation of these t's is given in Table XIII. Significance was determined by referral to the "Table of t" in Edwards (1955). The degrees of freedom, appropriate to this procedure, are equal to (n - 3). Since the null hypotheses required a two tailed test, and since the level of significance was set at .05, the table shows that with 1420 degrees of freedom, a "t" value equal to or greater than 1.96 is required in order that the null hypotheses be rejected.

TABLE XIII. Correlations Between Experimental Tests and Instructor Grades and Correlations Between Experimental Tests, Used in Computing "t's" for Significant Differences Between Validity Coefficients for Experimental Tests.

	r_{pi}	r_{ri}	r_{mi}	r_{pr}	r_{pm}	r_{mr}
NATURAL SCIENCE	.81	.75	.61	.77	.69	.66
SOCIAL SCIENCE	.65	.59	.49	.51	.53	.46

Inspection of the "t's" in Table XII reveals that all of the "t" values exceed the value of 1.96 necessary for rejecting the null hypotheses. Since the correlation coefficients descend in level of magnitude from the "Peaked Test" to the "Multimodal Test", and since all of the differences in the sizes of the validities are significant, it may be concluded that the "Peaked Test" is a better discriminator than either the "Rectangular Test" or the "Multimodal Test". It also follows that the "Rectangular Test" is better than the "Multimodal Test" in this respect.

A comparison of the validities of the experimental tests with those of the original tests should give some indication of these tests' relative ability to discriminate between examinees on the behavioral characteristic being measured. A statistical comparison between these validities is not possible since no test is known to the author which can be used to determine whether or not significant differences exist among them. In order to make the rational comparison as meaningful as possible the validities of the experimental tests were adjusted for length. This was accomplished by using the following formula, reported by Thorndike (1963).

$$r_{on} = \frac{r_{ol}}{\sqrt{\frac{1}{n} + (1 - \frac{1}{n})r_{11}}}$$

where r_{on} is the validity of the lengthened test,

r_{ol} is the validity of the original test,

r_{11} is the reliability of the original test, and,

n is the ratio of the length of the lengthened test to that of the original test.

The original Natural Science tests were each composed of 125 items. Since the experimental tests contained 50 items each, "n" was equated to 2.5 for computing the adjusted validity coefficients of the experimental test scores in this area.

In the Social Science area the three original tests contained 100, 120, and 110 items respectively, for an average of 110 items. "n" was, therefore, set equal to 2.2 for computing the adjusted validities of the Social Science experimental tests, since each contained 50 items.

The validities of the experimental tests corrected for length appear in Table XIV. The average validities of the original tests used to supply items for the experimental tests were also given in Table XIV.

TABLE XIV. Validity Coefficients for Original and Experimental Tests Adjusted for Length.

	Original Test	Peaked Test	Rectangular Test	Multimodal Test
NATURAL SCIENCE	.71	.86	.83	.65
SOCIAL SCIENCE	.62	.72	.64	.56

It is assumed that the items needed to increase the length of the experimental tests would be similar to the existing items of the experimental tests. The average reliability index for the items of the experimental tests would be larger than those of the original tests since items with a low index of discrimination were systematically eliminated from the experimental tests. According to Gulliksen (1962) this would tend to decrease the validity of the experimental tests since the validity of a test is equal to the ratio of its average validity index to its average reliability index. It can, therefore, be concluded from Table XIV that the validity coefficients for the "Peaked" and "Rectangular" tests are greater, in both areas, than are those of the original tests.

A Comparison of Group Means

As stated in Chapter I, one of the two methods by which validities were to be compared in this investigation is to determine which of three distributions of item difficulty, used in constructing an achievement test, will most effectively separate a group of examinees into subgroups for grading purposes. In order to answer this question the subjects used in this investigation were separated into criterion groups according to the sum of the numerical grades assigned by their instructors for the three terms being considered. Significance tests were performed to determine which of the experimental tests produced scores best able to discriminate between adjacent groups.

Students were assigned to criterion groups on the basis of total instructor grades as follows (see p. 18).

TOTAL INSTRUCTOR GRADE	CRITERION GROUP
38 thru 45	A
29 thru 37	B
20 thru 28	C
11 thru 19	D
3 thru 10	F

The results of this procedure are shown in Table XV. The numbers of individuals in each group are listed under N in this table. No "F" group appears in the table, since only one individual was assigned to this group and that was in the area of Social Science. This result was expected since students rarely continue through the entire three course sequence if they fail the first one or two courses of that sequence. Table XV also lists the criterion group score means for each of the experimental tests.

Figures 1 and 2 present these criterion group means in graphic form. The relative slopes of these lines as well as the relative slopes of the short segments connecting the mean score points of each group give an indication of the relative distances between the means. If it is assumed that group standard deviations, are not significantly different for each criterion group, then the slopes of these lines and also the slopes of the line segments should indicate the ability of the corresponding tests to

discriminate between groups.

TABLE XV. Numbers in Each Criterion Group, Groups Means, and Group Standard Deviation for the Experimental Tests in Natural Science and Social Science.

Criterion		PEAKED		RECTANGULAR		MULTIMODAL	
Groups	N	Mean	s.d.	Mean	s.d.	Mean	s.d.
			NATURAL	SCIENCE			
A	96	39.30	4.22	34.66	3.77	37.30	3.79
B	459	32.41	5.35	30.38	3.68	34.47	2.85
C	725	25.02	5.64	24.80	4.12	30.36	4.05
D	143	18.75	4.54	21.20	3.98	26.58	3.76
			SOCIAL	SCIENCE			
A	54	37.24	4.33	37.70	4.91	38.52	3.90
B	272	31.92	4.76	31.78	4.04	34.72	3.05
C	475	25.76	4.94	26.94	5.09	30.91	3.99
D	107	23.13	4.21	24.84	3.66	29.53	3.76

Inspection of Figure 1 reveals that the "Peaked Test" has both an over-all greater slope and also steeper line segments between each criterion group than the other experimental tests in the Natural Science area. These facts give tentative support to the hypothesis that the "Peaked Test" was the best discriminator between criterion groups.

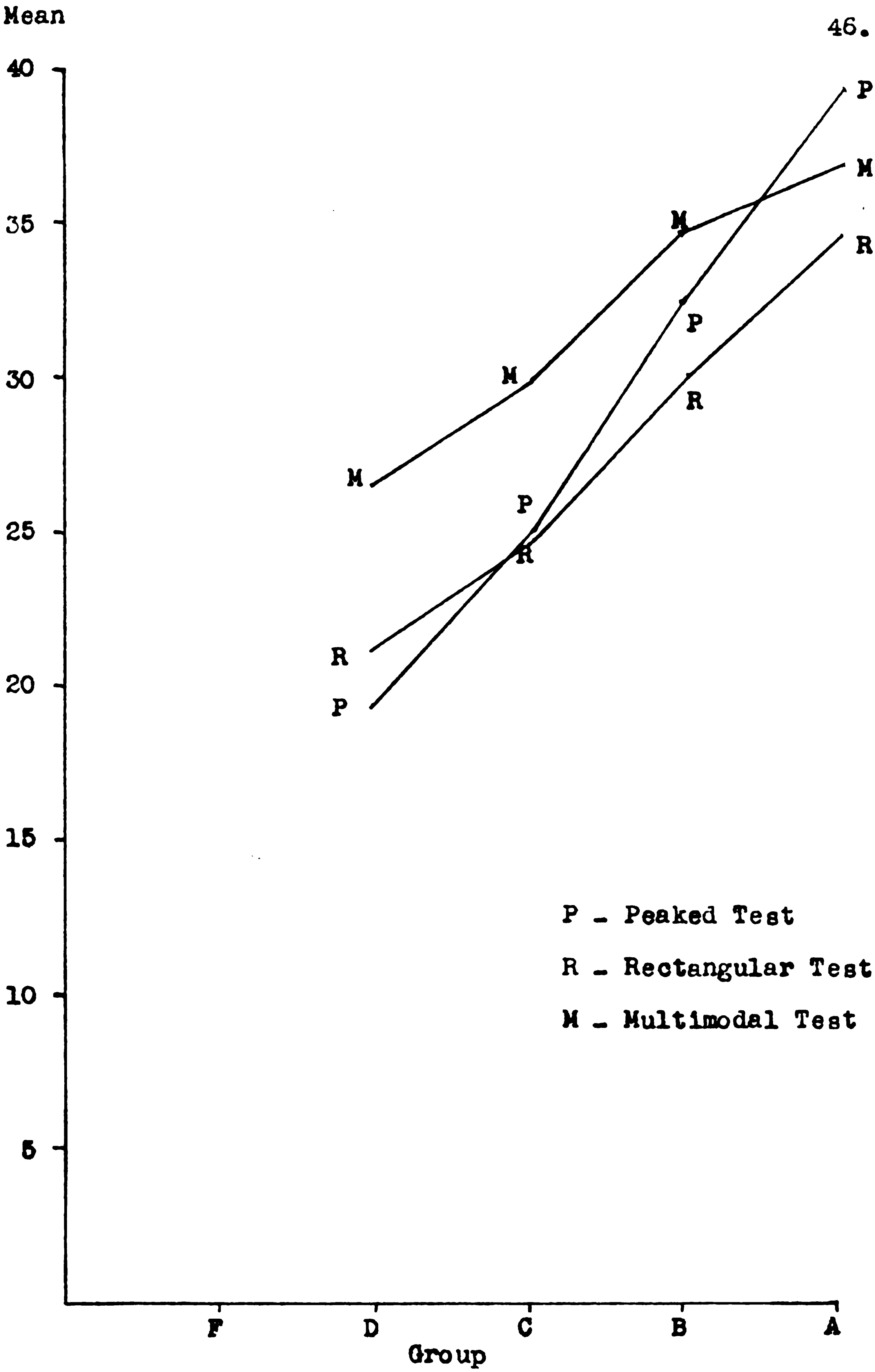

Fig. 1 Criterion Group Means For The Experimental Tests In Natural Science

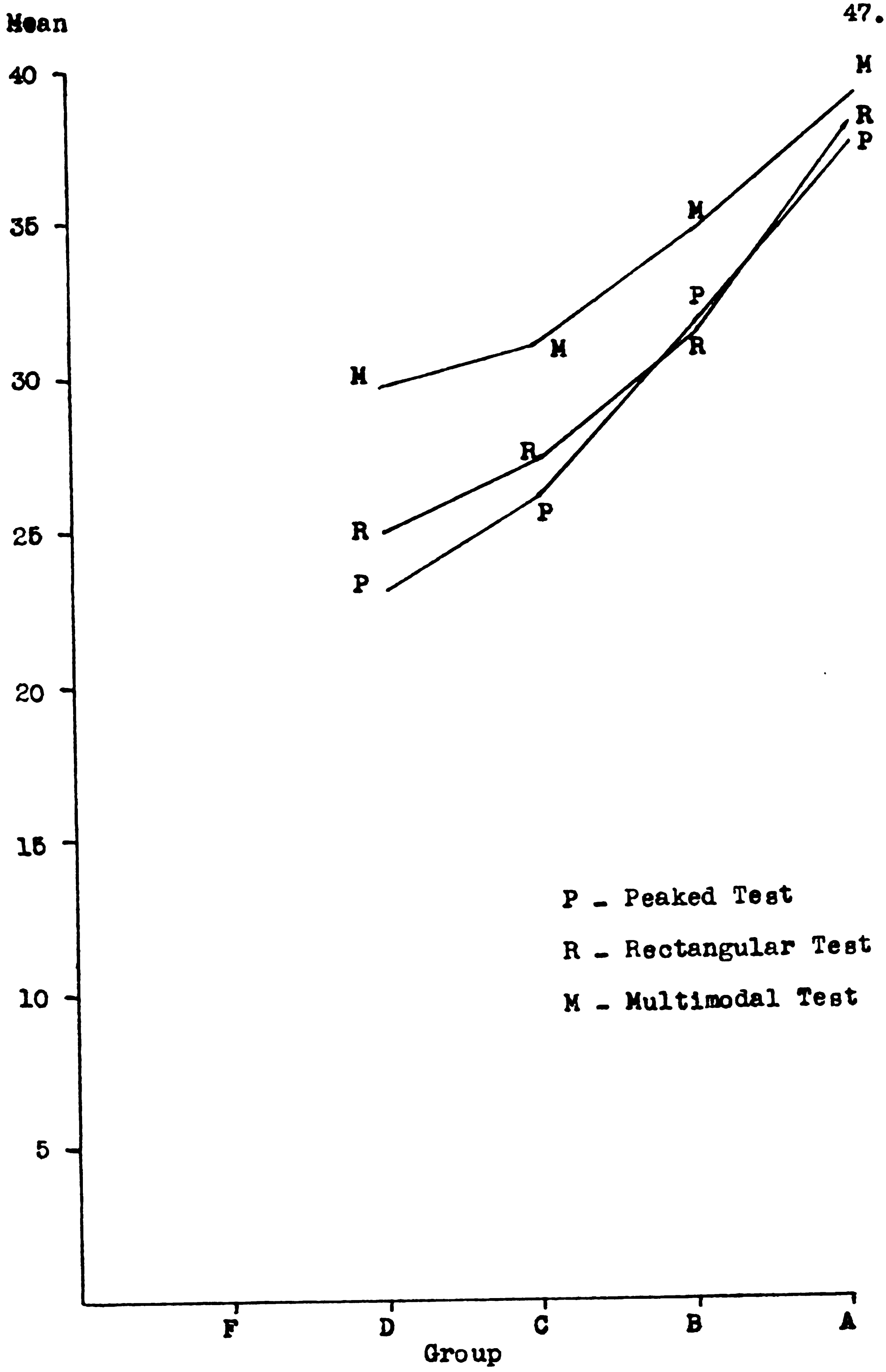

Fig. 2 Criterion Group Means For The Experimental Tests In Social Science

It is also of interest to note that the segment from "C" to "B" for the "Peaked Test" has a greater slope than any other segment on the chart. Examination of Table XV reveals that these two groups also have larger standard deviations than any of the other groups. Whether or not these larger dispersions of scores within the groups, and hence greater overlap between them, will seriously affect the ability of the test to discriminate between the criterion groups, can only be determined by a significance test. This test will follow this discussion of Figures 1 and 2.

Examination of Figure 2 reveals that apparently the "Peaked" and "Rectangular" tests in the Social Science area are both better discriminators among the criterion groups than the "Multimodal Test", since both have overall greater slopes. The "Peaked Test" seems to be better in differentiating "C's" from "B's", while the "Rectangular Test" seems to discriminate more effectively between group "A" and group "B". These observations can only be tentative since a statistical procedure using the variances of group test scores is necessary in order to determine whether or not these differences are actually significant. A test is also desirable in order to determine which differences are significant.

The statistical tests indicated above were performed in order to determine whether or not criterion group mean differences were significant. These differences between adjacent criterion group score means were computed and are listed in Table XVI. These differences were converted to z's by using the following formula:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{N_1 S_1^2 + N_2 S_2^2}{N_1 N_2}}}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the means of two groups,

N_1 and N_2 are the numbers of individuals in groups 1 and 2, and

S_1^2 and S_2^2 are the variances of the scores of groups 1 and 2.

The values for $\bar{X}_1 - \bar{X}_2$ were taken from Table XVI. Table XV lists the values for N. The values of S^2 can also be computed from the s.d.'s in this table. The values for the resulting z's are listed in Table XVII.

The null hypotheses, being tested, is that the population mean of a group's test scores is equal to the population mean of an adjacent group's test scores. If the level of significance is set at .05 and a one-tailed test

is performed, the table of the normal curve indicates that a "z" value equal to or larger than 1.65 will permit the rejection of the null hypotheses. An examination of the "z" values in Table XVII reveals that all of them are larger than this value. It may be concluded, therefore, that significant differences exist among the population means of adjacent letter group test scores, on all of the experimental tests.[1]

TABLE XVI. Differences Between the Means of Adjacent Criterion Groups on the Experimental Tests for Natural Science and Social Science.

Criterion Groups	Peaked Test	Rectangular Test	Multimodal Test
		NATURAL SCIENCE	
A - B	6.89	4.28	2.83
B - C	7.39	5.58	4.11
C - D	6.27	3.60	3.78
		SOCIAL SCIENCE	
A - B	5.32	5.92	3.80
B - C	6.16	4.84	3.81
C - D	6.27	3.60	3.78

[1] The more precise t-test could have been used in making these comparisons. The z's were used, however because they were the values that were computed in the following analysis, because the numbers of cases were generally large, and because the z's were so large that it was clear the more precise test would lead to the same conclusion.

TABLE XVII. z's for the Differences Between Criterion Groups for the Experimental Tests in Natural Science and Social Science.

Criterion Groups	Peaked Test	Rectangular Test	Multimodal Test
		NATURAL SCIENCE	
A - B	11.87	10.19	8.32
B - C	22.39	23.25	18.68
C - D	12.54	9.73	10.22
		SOCIAL SCIENCE	
A - B	7.60	9.40	7.92
B - C	16.65	13.44	13.61
C - D	12.06	6.92	9.00

Figures 3 and 4 are visual representations of the critical ratios for criterion group differences on each experimental test and were taken from Table XVII. The z's were plotted on the ordinate and assuming that all other group parameters were equal, the highest ordinate for each group difference indicated the corresponding test best able to discriminate between those particular groups. In keeping with this rationale, Figure 3 reveals that for the area of Natural Science, the "Rectangular Test" discriminated best between the "B" and "C" groups and that the "Peaked Test" discriminated most effectively, "A's" from "B's" and "C's" from "D's".

In the Social Science area, Figure 4 shows the "Peaked Test" to differentiate best between "C's" and "D's" and "B's" and "C's". The "Rectangular Test" discriminated best between group A and group B. These observations were based on the assumption that group sizes, variances and covariances were equal. Since this assumption was violated a statistical procedure was necessary in order to determine which experimental test discriminated most effectively between adjacent criterion groups.

A procedure for this purpose was developed by Saupe, (1965). He called it an "approximate, large sample test for comparing the ability of two measures to discriminate between two groups". This test may be expressed by the formula:

$$z_3 = \frac{z_1 - z_2}{\sqrt{2\left[1 - \frac{N_B C(X_{1A})(X_{2A}) + N_A C(X_{1B})(X_{2A})}{\sqrt{(N_A S^2_{1A} + N_B S^2_{1B})(N_A S^2_{2A} + N_B S^2_{2}B)}}\right]}}$$

where "z_1" and "z_2" are critical ratios of the difference to the standard error of the difference between the mean scores of two adjacent groups,

N is the number of individuals in each group,

S^2 is the variance of the scores in a group, and

$C(X_{1A})(X_{2A})$ is the covariance of the soores of two groups.

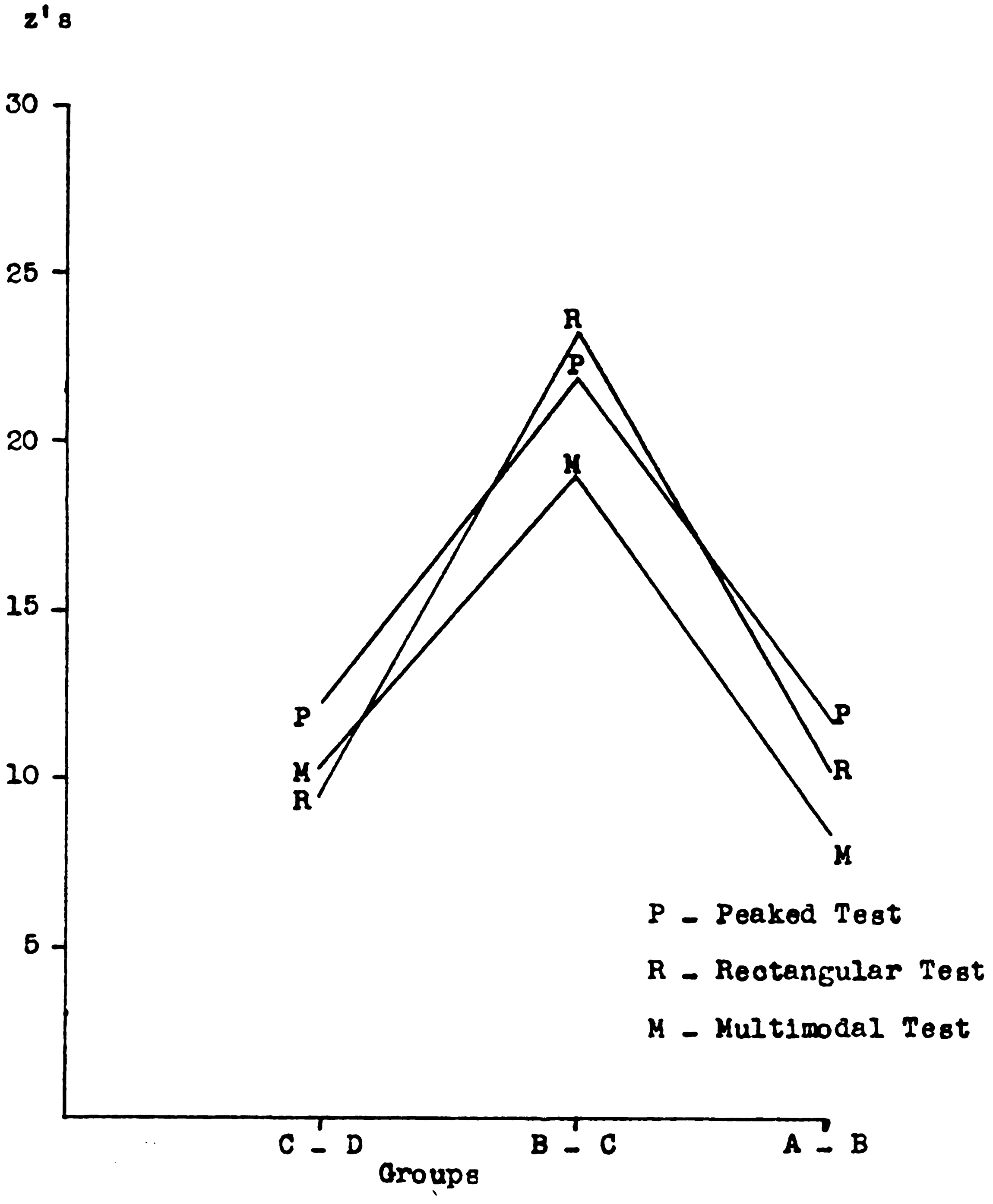

Fig. 3 z's Between Adjacent Criterion Groups For The Experimental Tests In Natural Science

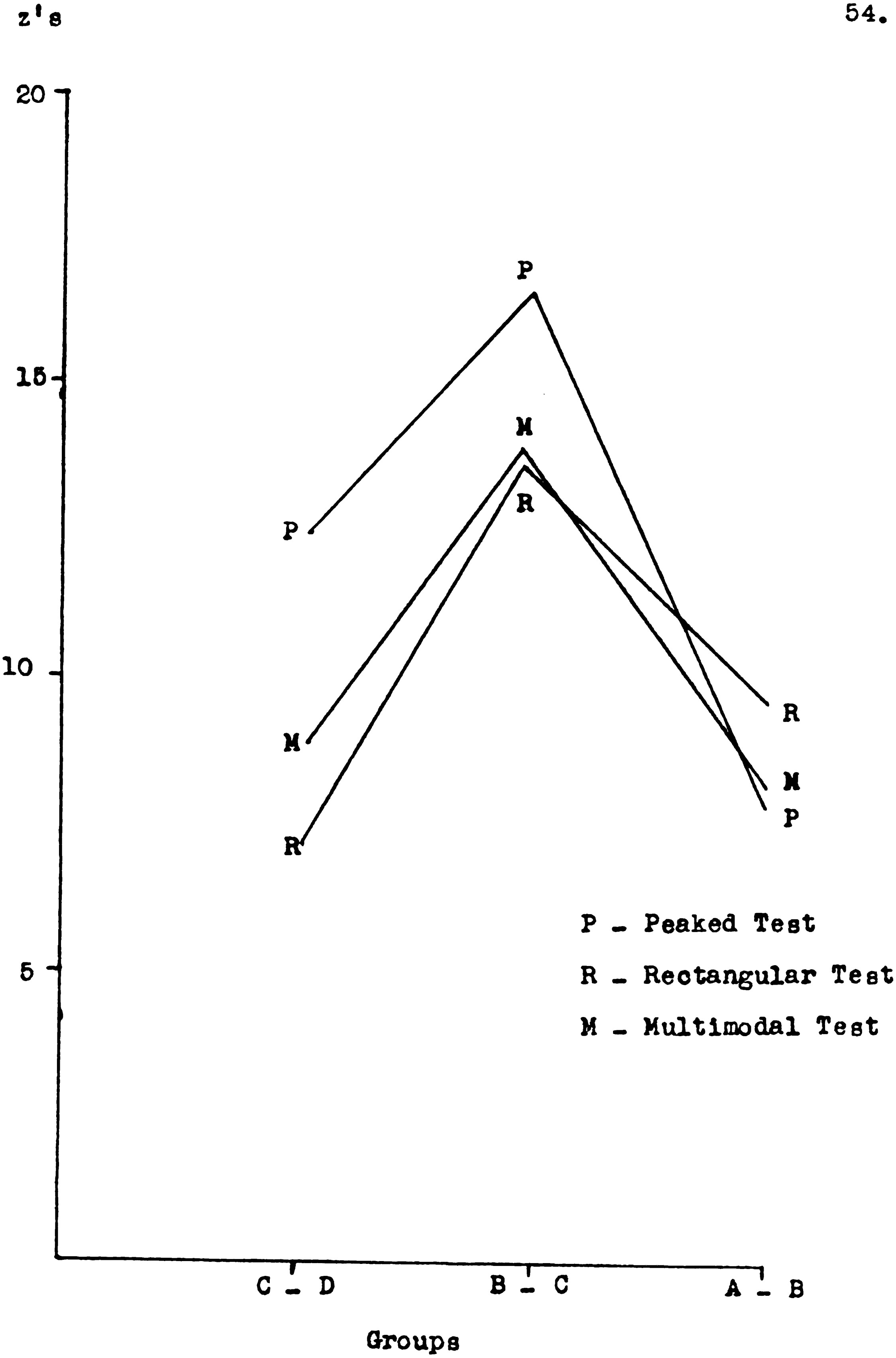

Fig. 4 z's Between Adjacent Criterion Groups For The Experimental Tests In Social Science

The values obtained from substituting the appropriate values in this formula are given in Table XVIII. Saupe (1965) has assumed that these z_3's values are normally distributed. A two-tailed test was used to test the null hypotheses that, the experimental tests in both subject areas, were equally effective in discriminating between adjacent criterion groups. The significance level was set at .05 and the table of the normal curve indicated that a value of 1.96 or larger, or -1.96 or smaller, was necessary in order to reject the null hypotheses. An asterisk appears above and to the right of the values in Table XVIII that are beyond these limits. It may be concluded that those values having an asterisk represent differences between critical ratios which are statistically significant. The corresponding experimental tests can be assumed to be the best discriminators for the groups and areas indicated.

An examination of Table XVIII shows that of the twelve comparisons of the "Peaked Test" with the other experimental tests, eight proved to be significantly in favor of the "Peaked Test". Of the four comparisons which were not significant, two were in favor of the "Rectangular Test" and one of these apprpached significance.

When compared only with the "Rectangular Test", three of the six comparisons were significantly in favor of the "Peaked Test". One other was in that direction, but not significant. The other two comparisons favored the "Rectangular Test" and one of these approached significance.

TABLE XVIII. "z '" Values for the Differences Between the 3 "z's" of Group Differences of the Experimental Tests in Natural Science and Social Science.

Criterion Groups	$z_p - z_r$	$z_p - z_m$	$z_r - z_m$
	NATURAL SCIENCE		
A's - B's	1.67	3.48*	1.46
B's - C's	-.89	3.28*	4.13*
C's - D's	2.40*	2.11*	-.46
	SOCIAL SCIENCE		
A's - B's	-1.94	-.26	1.96*
B's - C's	2.63*	2.53*	-.14
C's - D's	4.02*	2.59*	-6.06*

The "Rectangular Test" proved to be a better discriminator than the "Multimodal Test" in two cases. Only in one case was the "Multimodal Test" significantly better than the "Rectangular Test", and in no case was it statistically superior to the "Peaked Test".

The general pattern of the z's involving "Peaked Tests" in Table XVIII seems to support the findings of the

Cronbach, Warrington study (1952). They concluded that a "Peaked Test" should be more valid except for examinees of extremely high or low ability, assuming low interitem correlations. This would imply a rise in the ability of the "Peaked Tests" to discriminate most effectively among the middle groups. This ability is indicated by Table XVIII and Figure 4, for the "Peaked Tests", since the measures (z's) of ability to discriminate do rise for the middle instructor grade groups. Figure 3 failed to reveal this trend. No "F" groups were available for this study and therefore, it can only be inferred that if this theory can account for the results of the "Peaked Tests", then the z's for the "D - F" groups should decline in magnitude.

The only z involving a "Peaked Test" which tends to negate this theory is that for discriminating between the B and C groups in the Natural Science area. In this instance a greater actual difference occurred between the group mean scores of the "B - C" groups on the "Peaked Test" than on the "Rectangular Test". However, when these differences were converted to critical ratios, the critical ratio of the "Rectangular Test" was larger, although not significantly so.

An explanation for this phenomenon is indicated by the fact that the variances of these two groups of test scores are roughly twice as large for the "Peaked Test" as for the "Rectangular Test". Those for the "B" groups are 28.66 and 13.53, while those for the "C" groups are 31.88 and 17.04 for the "Peaked" and "Rectangular" tests respectively. These large group variances indicate a high degree of overlap for the scores of the adjacent criterion groups. This functioned to decrease significance of the difference between the means of these groups.

Since the test variance is equal to the square of the sum of the item reliability indices (Gulliksen 1962), the larger variances for groups B and C imply that for these groups in the Natural Science area the items of the "Peaked Test" had relatively large interitem correlations. This increase in the homogeneity of the items also indicated that the "Peaked Test" became much more reliable for the two groups with the result that there was an accompanying decrease in the validity for these criterion groups. The explanation for this apparent paradox is that in practice as the reliability of a test increases the validity also increases up to a certain point and then as reliability continues to increase, validity decreases. Since validity is usually computed from a complex criterion, it should not appear strange that a test having a high degree of item homogeneity should be a poor predictor for a criterion heterogeneous in nature. In reality, as a test becomes more reliable the specificity of measurement increases and usually the number of factors being measured decreases. If the validating criterion consists of only those factors being measured by the test, then an increase in reliability could be expected to bring about an accompanying increase in validity. On the other hand, if an increase in reliability results in measuring fewer of the factors relevant to the validating criterion, an improvement in validity can not be expected. Apparently this latter case is what happened to the "Peaked Test" for the "C" and "D" groups in the Natural Science area.

The Cronbach and Warrington position has taken this phenomenon into account. It advocated a widening of the range of item difficulty when high correlations existed among items. This position would, therefore, account for the lack of validity for the "Peaked Test" in the Natural Science area, for the "B" and "C" criterion groups, on the basis that the high interitem correlations for these groups required a greater spread in item difficulty.

This position would also explain the greater validity of the "Rectangular Test" for these groups in this area. If it can be assumed that the interitem correlation for this test were similar to those of the "Peaked Test", then the greater dispersion of item difficulty in the "Rectangular Test" would be expected to produce greater validity.

The evidence in this section indicated that the "Peaked Test" was more effective in discriminating among criterion groups than the "Rectangular Test" in three of the four comparisons involving groups in the middle range of ability. It was assumed that these results were due to moderate interitem correlations for the criterion groups involved. There was some indication that the "Rectangular Test" was a better discriminator between adjacent criterion groups than the "Peaked Test" for the other comparison involving groups in the middle range of ability. This result was assumed to be a reflection of high interitem correlations for the groups involved. Of the two comparisons of groups available at the extremes of ability, one favored the "Peaked Test" while the other favored the "Rectangular Test". Neither comparison revealed a significant difference. The "Multimodal Tests" apparently were the poorest discriminators of the experimental tests.

Summary

In this phase of the investigation data gathered from the experiment were presented and analyzed. The results were compared statistically where possible and rationally where no statistical test was available.

A rational examination of the reliability coefficients of the experimental test indicated that the "Peaked

Test" was most reliable in the Natural Science area. No real difference between the reliabilities of the "Peaked Test" and the "Rectangular Test" was apparent in the area of Social Science. The "Multimodal Test" apparently was the poorest of the three experimental tests in regard to reliability. The "Peaked Test" had an adjusted reliability coefficient larger than that of the original test in the area of Natural Science. Both the "Peaked" and the "Rectangular" tests had larger adjusted reliabilities in the Social Science area than the original tests.

Statistical evidence indicated that the "Peaked Test" produced the highest validity coefficients of the experimental tests when instructor grades were used as the validating criterion. The "Rectangular Test" was next, and the "Multimodal Test" was last in this respect. The validities of the experimental tests were adjusted for length, and it was assumed that items added would have characteristics similar to those of the existing items. These adjusted validities of both the "Peaked" and the "Rectangular" tests exceeded the validities of the original tests in both subject areas.

When instructor-grade groups were used as the validating criteria, the "Peaked Test" was found to be the best discriminator between most of the adjacent criterion groups in the middle of the ability range. Groups were available only for the upper extremes of ability and comparisons between test score means reveals one critical ratio in favor of "Peaked Tests" and one in favor of the "Rectangular Test". The "Multimodal Test" apparently was the poorest discriminator of the experimental tests.

CHAPTER IV

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Summary

The purpose of this investigation was to determine which of three distributions of item difficulty, used to construct achievement tests, would be most effective in terms of the homogeneity of their scores and their validities for grading purposes.

Achievement test data for two subject areas was available at Michigan State University. These data were used to investigate the problem. The items from three of these tests, given in sequence to the same students, were combined to form an item pool for each of the two subject areas. Item analyses were available for these tests and provided item difficulties and discrimination indices, which were used in the construction of 50-item experimental tests.

A "Peaked Test" in which item difficulty ranged from .46 to .63, a "Rectangular Test" with a range in item difficulty from .22 to .93, and a "Multimodal Test" with items concentrated at four levels of difficulty, were constructed for each subject area.

Reliability coefficients were determined for each test by use of Kuder Richardson #20 formula. These coefficients were compared with each other rationally. The "Peaked Test" was found to be most reliable in the Natural Science area. No real difference was apparent between the reliabilities of the "Peaked" and the "Rectangular" tests in the Social Science area, although both had higher reliability coefficients than that of the "Multimodal" test.

The reliabilities of the experimental tests were corrected for length and compared with those of the original tests. It was acknowledged that items with low discrimination indices on the original tests were not included in the experimental tests and that this would tend to increase the reliabilities of the experimental tests. Only the "Peaked Test" had an adjusted reliability coefficient higher than that of the original test in the Natural Science area. Both the "Peaked" and the "Rectangular" tests were more reliable in the Social Science area. The "Rectangular Test" in this area had a reliability coefficient only slightly larger than that of the "Peaked Test".

The validity of the experimental tests was determined in two ways. First, the total test score for each of the experimental tests was correlated with a criterion which consisted of the total instructor grade for a three course sequence. The resulting validity coefficients were compared statistically with each other and were also adjusted for length to correspond with the lengths of the original tests and then compared rationally. The statistical comparisons showed that the "Peaked Tests" produced higher validity coefficients than any of the other experimental tests. The "Rectangular Test" was superior to the "Multimodal Test" in this respect.

Rational comparisons of adjusted experimental test validities with original test validities assumed that the items added to make the experimental tests as long as the original tests were similar to the existing items, and that items with low discrimination indices eliminated from the experimental tests would tend to decrease their validity. It can therefore be concluded that validity coefficients for the "Peaked" and "Rectangular" tests were greater in both areas than were those of the original tests.

The second means of determining validity was to compare the relative abilities of the experimental tests to discriminate between adjacent instructor grade groups. This was accomplished by computing critical ratios for the differences between adjacent grade groups for each experimental test. A statistical comparison of these critical ratios indicated that in a majority of the comparisons the "Peaked Test" discriminated most effectively between adjacent criterion groups in the middle of the range of ability. Of two comparisons for the upper extremes of ability, one favored the "Peaked Test" and one favored the "Rectangular Test". The "Multimodal Test" apparently was the poorest discriminator of the experimental tests.

Limitations

The main purpose of this study was to determine by empirical means which of three distributions of item difficulty, used in the construction of achievement examinations, would result in the most useful instruments for grading purposes. Conclusions and recommendations from this study are tempered by a number of limitations which should be pointed out.

This investigation was performed with data available on academic achievement tests in courses at the college level. This data was compiled under practical conditions as they existed in the actual construction of college achievement examinations for large numbers of students. Item statistics were estimated from a sample taken from the tails of the distribution of test scores. While this procedure provides statistics which can be used in practical situations with some justification, their use in this study necessitates the limitation of the inferences of the results to similar situations.

The construction of 50-item experimental achievement tests, whose items had the desired characteristics, required a large supply of items that could be accumulated only by the pooling of items from three examinations. This procedure resulted in the attrition of students of low ability and resulted in the complete loss of an "F" instructor grade group. Some students of very high ability also were lost to the study. The assumption was made that the ability levels of the groups were not greatly affected by these losses. However, lack of statistical evidence of this fact also limits the inferences which can be made from the results of this study, since item statistics for each course were computed for the different groups rather than for a total group which could be shown to have an ability level comparable to that of the experimental group.

The pooling of results from three courses also resulted in the combining of instructor grades. In some cases students may have been taught by the same instructor for all three terms, and in other cases a different teacher may have conducted a student's class for each term. The large numbers of students involved in these three courses required the services of a large number of instructors whose subjective evaluations and personalities certainly had some influence on the grades which they assigned. Whether or not this had some biasing effect on the actual instructor grades used in this investigation is not known.

Comparison of test results was hampered in the case of the reliability coefficients since there was a lack of appropriate statistical tests which would lead to more definitive conclusions regarding these results. The comparison of adjacent grade group means was made by an approximate test which makes acceptance of the results somewhat tentative in nature.

A further limitation was the lack of enough items with the precise characteristics desired in order to construct each experimental test in complete accordance with its respective theory. However, in practice, this condition also exists since it is difficult to obtain a plentiful supply of useful items.

Conclusions

The conclusion of this study are dependent upon the assumption that item characteristics other than item difficulty such as, discrimination indices, reliability indices, and validity indices were the same for the different types of tests.

In so far as the techniques employed in this investigation may be justified, the following conclusions seem defensible:

1. College achievement tests which have items concentrated within a small range of difficulty, somewhere near the mean ability level of the group, have a tendency to produce larger reliability coefficients than either

tests which have items covering a wide range of difficulty levels, or tests whose items are concentrated at four different ability levels. A major factor accounting for this is the greater item variance for items near the mean ability level of the group. This results in a larger item standard deviation and hence a greater reliability index, provided that discrimination indices are held constant, and the result is a larger reliability coefficient.

2. College achievement tests which have a wide range of item difficulties have a tendency to be more reliable than those whose items are concentrated at four different difficulty levels not including the middle range. This conclusion results from the fact that omission of items near the mean level of difficulty tends to decrease item reliability indices which are dependent upon item standard deviations as well as discrimination indices. If item discrimination indices are held constant a decrease in item standard deviation results in a decline in the reliability index, and hence, results in a smaller reliability coefficient.

3. Validity coefficients for college achievement tests (correlations with independently assigned instructor grades) will be larger for tests whose items are concentrated near the level of mean group ability, than for those tests whose items are concentrated at four different levels,

or for tests with a wide range of item difficulty when the validating criterion is instructor grades. This conclusion assumes that interitem correlations are similar to those used in this study.

4. College achievement tests composed of a wide range of item difficulties will correlate higher with instructor grades than will tests whose item difficulties are concentrated at four different levels assuming that interitem correlations are similar to those used in this investigation.

5. College achievement tests with a small range of item difficulties concentrated near the average ability level of the examinees and with moderate interitem correlations have a tendency to discriminate more effectively among instructor grade groups than those tests whose item difficulties have either a wide range or are concentrated at four levels. This is especially true for the groups in the middle ranges of ability and is dependent upon interitem correlation being similar to those used in this study in the area of Social Science.

6. College achievement tests having a relatively high degree of interitem correlations will discriminate more effectively when the variance of item difficulties is greater than when items are concentrated near the mean level of difficulty. Test validity is a function of the

sum of the variance of item unreliability and the variance of item difficulties. Validity is maximum for one score when this sum is small, but validity increases for a wider range of scores as this sum of variances increases, up to a certain point. Therefore, when interitem correlations are high (low unreliability) a compensatingly larger variance of item difficulty is necessary in order to increase the sum of variances and hence improve the validity of the test for a number of scores.

7. College achievement tests whose item difficulties cover a wide range, discriminate more effectively among groups than those whose items are concentrated at four ability levels other than near the mean level of ability. This results from the fact that validity is dependent upon the ratio of average validity index to average reliability index. Since the average reliability indices of the two tests are assumed to be equal, the magnitudes of the validities of the tests are dependent upon the corresponding sizes of the average validity indices. The major discernable difference between the two tests is the inclusion of items in the middle range of ability on one test and their omission on the other test. It is, therefore, concluded that the test containing items near the mean level of difficulty has a higher validity due to the influence of these items in increasing the average validity index.

Recommendations

In so far as the techniques employed in this study may be valid, the following recommendations seem justified:

1. If it can be assumed that items available for achievement tests have item characteristics similar to the items used in this investigation, it is recommended that in the construction of achievement tests as many of the items as possible be located near the mean ability level of the examinees.

2. The validities of the "Peaked Tests" were clearly superior when individual test scores were correlated with instructor grades. The results of comparing criterion group means presented some evidence that they were also the best discriminators among some of these groups. There was also an indication that the "Rectangular Tests" were somewhat effective in this respect. A study is, therefore, recommended which would determine empirically whether or not a distribution of item difficulties intermediate between "Peaked" and "Rectangular" and "Normal" in shape, would be more effective than either the "Peaked" or the "Rectangular" distribution in discriminating among the whole range of criterion groups.

3. This study was designed to investigate the relationship which exists between item difficulties and

the criteria which instructors use for grading purposes. The assumption was made that item difficulties were not related to item content. The extent to which this assumption is invalid will affect the accuracy of the results of this investigation. It is, therefore, recommended that a study be undertaken to investigate the extent and nature of any relationship which may exist between item content and item difficulty.

BIBLIOGRAPHY

Brogden, H. E. "Variations in Test Validity with Variation in the Distribution of Item Difficulties, Number of Items, and Degree of Their Intercorrelation." *Psychometrika*, 1946, *XI*, 197-214.

Brogden, H. E. "On the Interpretation of the Correlation Coefficient as a Measure of Predictive Efficiency." *Journal of Educational Psychology* 1946, *37*, 65-76.

Committee on Measurement and Evaluation, *College Testing*, Washington: American Council on Education, 1949, p. 40.

Cronbach, L. J. and Warrington, W. W. "Efficiency of Multiple-Choice Tests as a Function of Spread of Item Difficulties." *Psychometrika*, June 1952, *17*, No. 2, 127-147.

Cronbach, L. J. *Essentials of Psychological Testing*, New York: Harper and Brothers., 1960, p. 131.

Cook, W. W. "The Functions of Measurement in the Facilitation of Learning." In E. F. Lindquist (ed.) *Educational Measurement*, Washington: American Council on Education, 1963, 3-46.

Davis, F. B. "Item Selection Techniques." In E. F. Lindquist (ed.) *Educational Measurement*, Washington: American Council on Education, 1963, 119-158.

Dressel, P. L. *Evaluation in Higher Education*, Boston: Houghton Mifflin Co., 1961.

Edwards, A. L. *Statistical Methods for the Behavioral Sciences*, New York: Rinehart and Co. 1955, p. 501.

Flanagan, J. C. "A Table of the Values of the Product Moment Coefficient of Correlation in a Normal Bivariate Population Corresponding to Given Proportions of Successes."

Flanagan, J. C. "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product Moment Coefficient From Data at the Tails of the Distribution." Journal of Educational Psychology, 1939, XXX 674-680.

Ebel, R. E. "Procedures for the Analysis of Classroom Tests." Educational and Psychological Measurement, 1954, 14, p. 359.

Ebel, R. E. "Item Difficulty and Test Effectiveness." Presented to the AERA, Atlantic City, Feb. 1959.

Gulliksen, H. "The Relation of Item Difficulty and Interitem Correlation to Test Variance and Reliability." Psychometrika, 1945, X, 79-91.

Gulliksen, H. Theory of Mental Tests, New York: John Wiley and Sons, Inc., 1962, p. 223.

Hayes, W. L. Statistics for Psychologists, New York: Holt, Rinehart and Winston, 1963, p. 500-501.

Hull, C. L. Aptitude Testing, Yonkers-on-Hudson: World Book, 1928.

Jackson, R. A. "An Item Analysis Technique Based Upon Adjacent Group Differences." Unpublished Ed.D. Thesis, Michigan State University, East Lansing, Mich. 1952.

Kelly, T. L. Statistical Method, New York: Macmillan, 1923.

Kelly, T. L. "The Selection of Upper and Lower Groups for the Validation of Test Items." Journal of Educational Psychology, 1939; 30, 17-24.

Lindquist, E. F. Statistical Analysis in Educational Research, New York: Houghton Mifflin Co., 1940

Lindquist, E. F. "Preliminary Considerations in Objective Test Construction". Educational Measurement, Washington: American Council on Education, 1963, 119-158.

Lord, F. M. "The Relation of the Reliability of Multiple Choice Tests to the Distribution of Item Difficulties." Psychometrika, 1952, XVII, 181-194.

Myers, C. T. "The Relationship Between Item Difficulty and Test Validity and Reliability." Educational and Psychological Measurement. XXII, No. 3, 1962, 565-571.

Noll, V. Introduction to Educational Measurement, Boston: Houghton Mifflin Co., 1957, p. 150.

Richardson, M. W. "Relation Between the Difficulty and the Differential Validity of a Test." Psychometrika, I, 1936, 33-49.

Saupe, J. "A Significance Test for Comparing the Ability of Two Measures to Discriminate Between Two Groups." Unpublished paper. East Lansing, Mich., 1965.

Saupe, J. "Technical Considerations in Measurement" In P. L. Dressel, (ed.), Evaluation in Higher Education, Boston: Houghton Mifflin Co., 1961, p. 444.

Thurstone, T. G. "The Difficulty of a Test and Its Diagnostic Value." Journal of Educational Psychology, 1932, XXIII, 335-343.

Thorndike, R. L. "Reliability" In Lindquist, (ed.) Educational Measurement, Washington: American Councel on Education, 1963, p. 608.

Traxler, A. E. Measurement and Evaluation in The Improvement of Education, Washington: American Council on Education, 1951.

Tucker, L. R. "A Note on the Estimation of Test Reliability by the Kuder-Richardson Formula #20." Psychometrika, XIV, 1949, 117-119.

Tyler, R. W. *Constructing Achievement Tests*, Columbus: Ohio State University, 1934.

_____Annual Report, Princeton: Educational Testing Service, 1963

_____"Item Analysis Summary of University College Examinations." Unpublished Bulletin. East Lansing, Mich: Office of Evaluation Services, Michigan State University, July 1, 1964.

APPENDIX

NATURAL SCIENCE
Peaked Test

ITEM NO.		DIFF.	DISC.	ITEM NO.		DIFF.	DISC.
1.	S105	.46	.45	26.	S26	.56	.60
2.	S54	.46	.52	27.	W102	.56	.76
3.	W2	.46	.49	28.	F10	.56	.49
4.	F69	.47	.58	29.	W59	.57	.47
5.	W24	.46	.33	30.	F75	.57	.43
6.	S103	.47	.58	31.	S92	.57	.47
7.	W44	.49	.57	32.	W30	.59	.44
8.	S104	.50	.44	33.	F33	.59	.48
9.	F3	.50	.37	34.	F43	.59	.44
10.	S79	.51	.27	35.	W96	.59	.72
11.	F34	.51	.50	36.	F82	.60	.66
12.	F91	.52	.40	37.	W115	.60	.63
13.	F99	.52	.33	38.	W117	.60	.58
14.	W16	.52	.29	39.	S88	.60	.58
15.	S63	.52	.33	40.	S48	.60	.66
16.	S74	.52	.48	41.	F80	.60	.54
17.	S89	.53	.42	42.	S87	.61	.61
18.	W32	.53	.42	43.	W99	.61	.53
19.	S60	.53	.42	44.	W116	.61	.61
20.	S24	.54	.53	45.	W98	.62	.55
21.	F2	.54	.41	46.	S73	.63	.67
22.	F56	.54	.40	47.	S44	.63	.58
23.	F111	.54	.33	48.	W13	.65	.38
24.	F90	.55	.43	49.	W79	.64	.39
25.	F87	.55	.35	50.	F104	.63	.63

NATURAL SCIENCE
Rectangular Test

ITEM NO.		DIFF.	DISC.	ITEM NO.		DIFF.	DISC.
1.	F109	.23	.25	26.	F5	.56	.53
2.	S68	.17	.65	27.	S53	.57	.39
3.	F51	.20	.30	28.	W109	.58	.69
4.	W41	.21	.53	29.	F107	.59	.44
5.	F78	.22	.40	30.	S121	.61	.45
6.	S106	.25	.34	31.	W123	.63	.33
7.	W71	.27	.28	32.	F54	.65	.47
8.	F63	.29	.41	33.	W54	.67	.43
9.	S64	.30	.33	34.	W97	.69	.66
10.	W86	.31	.35	35.	F77	.72	.45
11.	S83	.33	.53	36.	S12	.72	.70
12.	F74	.35	.42	37.	W51	.74	.73
13.	S99	.37	.37	38.	S17	.75	.58
14.	F81	.39	.40	39.	S7	.77	.56
15.	F70	.41	.36	40.	W106	.79	.45
16.	W34	.43	.36	41.	W6	.81	.59
17.	F85	.44	.45	42.	F37	.83	.56
18.	S31	.45	.51	43.	W27	.85	.53
19.	F92	.46	.45	44.	S70	.84	.63
20.	F31	.46	.41	45.	W12	.86	.40
21.	S97	.47	.31	46.	S51	.87	.50
22.	W82	.49	.38	47.	S61	.88	.48
23.	F105	.52	.67	48.	S122	.89	.46
24.	S59	.54	.33	49.	W93	.90	.30
25.	W18	.55	.51	50.	S114	.93	.25

NATURAL SCIENCE
Multimodal Test

ITEM NO.	DIFF.	DISC.	ITEM NO.	DIFF.	DISC.
1. F114	.12	.40	26. S3	.78	.34
2. S82	.12	.48	27. W104	.79	.62
3. W60	.14	.31	28. F112	.79	.38
4. F42	.16	.36	29. S96	.80	.51
5. F113	.20	.51	30. W114	.80	.61
6. F95	.20	.44	31. W53	.80	.44
7. W101	.41	.81	32. W15	.80	.37
8. W55	.42	.37	33. F67	.80	.68
9. F16	.42	.25	34. W118	.81	.59
10. F47	.43	.35	35. S9	.81	.42
11. S19	.43	.51	36. S11	.81	.42
12. S32	.43	.31	37. S36	.81	.35
13. S95	.43	.47	38. F25	.82	.32
14. S109	.43	.43	39. F50	.82	.58
15. S98	.44	.37	40. W63	.82	.58
16. S71	.44	.60	41. W107	.82	.40
17. S22	.44	.53	42. W14	.89	.33
18. S18	.44	.49	43. W93	.90	.30
19. F9	.45	.43	44. S69	.91	.51
20. F106	.45	.54	45. F21	.92	.37
21. F117	.45	.66	46. W42	.92	.49
22. F44	.78	.34	47. S52	.95	.25
23. W49	.78	.40	48. S90	.95	.46
24. W111	.78	.40	49. F38	.95	.25
25. S93	.78	.70	50. W10	.97	.30

SOCIAL SCIENCE
Peaked Test

ITEM NO.	DIFF.	DISC.	ITEM NO.	DIFF.	DISC.
1. S110	.48	.25	26. F28	.56	.41
2. S59	.48	.29	27. W49	.56	.33
3. F1	.48	.29	28. W24	.56	.68
4. F68	.49	.31	29. S13	.56	.33
5. S92	.49	.46	30. S101	.57	.51
6. F82	.50	.33	31. F50	.57	.43
7. W8	.50	.29	32. S47	.58	.33
8. S104	.50	.56	33. S15	.58	.53
9. S12	.50	.44	34. W105	.58	.33
10. S61	.51	.31	35. W46	.58	.46
11. S7	.51	.22	36. F29	.58	.53
12. W108	.51	.35	37. F23	.59	.44
13. F10	.51	.22	38. W27	.59	.32
14. S69	.53	.35	39. S76	.60	.38
15. W55	.53	.39	40. F81	.60	.30
16. W45	.53	.36	41. F39	.60	.21
17. W80	.53	.36	42. F30	.61	.49
18. F80	.54	.33	43. S94	.61	.45
19. W2	.54	.41	44. S25	.61	.36
20. W83	.54	.25	45. F26	.61	.65
21. S4	.54	.33	46. W103	.62	.39
22. S84	.54	.67	47. W102	.62	.43
23. S1	.55	.23	48. F63	.63	.54
24. W16	.55	.23	49. F22	.63	.41
25. W99	.55	.35	50. F27	.63	.41

SOCIAL SCIENCE
Rectangular Test

ITEM NO.		DIFF.	DISC.
1.	S83	.22	.22
2.	S20	.24	.44
3.	W61	.29	.31
4.	S40	.30	.38
5.	W59	.31	.55
6.	F45	.32	.56
7.	S16	.33	.20
8.	S67	.34	.36
9.	W101	.36	.31
10.	S37	.38	.26
11.	F49	.39	.32
12.	S100	.40	.21
13.	F31	.41	.36
14.	F9	.44	.60
15.	W17	.45	.35
16.	W79	.46	.60
17.	F92	.47	.31
18.	W9	.48	.33
19.	S2	.49	.46
20.	W65	.50	.29
21.	F2	.51	.38
22.	S79	.53	.42
23.	S28	.54	.21
24.	W89	.55	.43
25.	W77	.56	.33
26.	F78	.57	.23
27.	F24	.58	.21
28.	S64	.59	.23
29.	F67	.60	.21
30.	F20	.63	.33
31.	W85	.65	.42
32.	F11	.66	.49
33.	F6	.67	.53
34.	F88	.68	.56
35.	W53	.70	.43
36.	S85	.71	.52
37.	S18	.73	.56
38.	S102	.74	.36
39.	F57	.75	.45
40.	S120	.76	.50
41.	F14	.77	.30
42.	S66	.80	.24
43.	W38	.83	.46
44.	F51	.85	.62
45.	S107	.86	.51
46.	F41	.87	.59
47.	W63	.88	.57
48.	S21	.89	.55
49.	W7	.91	.51
50.	W12	.92	.23

SOCIAL SCIENCE
Multimodal Test

ITEM NO.		DIFF.	DISC.	ITEM NO.		DIFF.	DISC.
1.	F75	.08	.23	26.	W88	.76	.50
2.	S41	.13	.38	27.	F99	.76	.32
3.	S30	.14	.31	28.	F87	.77	.42
4.	S42	.16	.21	29.	F59	.77	.64
5.	S45	.22	.34	30.	F13	.77	.36
6.	F65	.35	.33	31.	S36	.77	.36
7.	F83	.35	.42	32.	S34	.77	.36
8.	S44	.35	.47	33.	S97	.78	.47
9.	S9	.36	.22	34.	S86	.78	.40
10.	F43	.36	.39	35.	S31	.78	.40
11.	W56	.36	.48	36.	W29	.78	.28
12.	W88	.38	.30	37.	W73	.78	.34
13.	W58	.38	.34	38.	F35	.78	.28
14.	S111	.40	.42	39.	F5	.78	.40
15.	F47	.40	.54	40.	W22	.79	.45
16.	W43	.40	.30	41.	F42	.91	.51
17.	W54	.40	.34	42.	F44	.91	.26
18.	W6	.40	.26	43.	S99	.92	.37
19.	S24	.75	.45	44.	S60	.93	.34
20.	S48	.76	.38	45.	F36	.93	.34
21.	S35	.76	.21	46.	F32	.93	.46
22.	S26	.76	.50	47.	S73	.94	.43
23.	W13	.76	.44	48.	W50	.95	.40
24.	W48	.76	.32	49.	W5	.95	.25
25.	W62	.76	.27	50.	F7	.95	.25