This dissertation has been 63-2477
microfilmed exactly as received

BRENNER, Marshall Hallock, 1933-
TEST DIFFICULTY, RELIABILITY, AND DISCRIM-
INATION AS FUNCTIONS OF ITEM DIFFICULTY
ORDER.

The Ohio State University, Ph.D., 1962
Psychology, general

University Microfilms, Inc., Ann Arbor, Michigan

TEST DIFFICULTY, RELIABILITY, AND DISCRIMINATION
AS FUNCTIONS OF ITEM DIFFICULTY ORDER

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Marshall Hallock Brenner, A.B., M.A.

* * * * *

The Ohio State University
1962

Approved by

Adviser
Department of Psychology

ACKNOWLEDGMENTS

The investigator is indebted to many persons for their aid in the completion of this study.

Invaluable guidance and counsel were received from the three members of the candidate's committee: Dr. Robert J. Wherry, who was the major adviser, Dr. John Horrocks, and Dr. Edward Furst.

The study could not have proceeded as smoothly as it did without the unselfish aid and cooperation of Dr. Horrocks, professor in charge of the educational psychology program. Dr. Horrocks helped solve numerous problems related to the administration of the tests.

The investigator is certainly appreciative of the outstanding manner in which all of the 407 instructors accepted the study and endeavored to help in its successful completion. Of special mention here are Mr. William Bricker, the assistant instructor for 407, and Mr. Robert Weatherford, who uncomplainingly bore the greatest load in helping the experimenter analyze, revise, and select items for the tests.

The analysis of data was completed with the aid of Mr. Robert Oates and his staff at the Orientation Program Office. Help in analysis was also received from Mr. Dickson Call at the Numerical Computation laboratory.

The more than twelve hundred students who were involved received a hearty word of appreciation, although most will never know.

And last, but always first, I thank my wife many times over for listening to, and encouraging, me through many weeks of single-minded discussions.

CONTENTS

	Page
ACKNOWLEDGMENTS	11
TABLES	v
Chapter	
I. THE PROBLEM	1
II. PROCEDURE AND RESULTS	9
General Remarks	9
Preliminary Procedure	11
Experiment I	15
Experiment II	21
Experiment III	31
Experiment IV	40
III. SUMMARY AND INTERPRETATION OF RESULTS	55
APPENDIX	
A. Difficulty and Discrimination Values for the the Items in the Quizzes Used in this Study	64
B. Test Statistics for Quizzes 1,2,3, and 4	84
C. Means of Students' Estimates of Difficulty for the Items in Quizzes 7	86
BIBLIOGRAPHY	89
AUTOBIOGRAPHY	92

TABLES

Table	Page
1. Number and Percentage of 407 Students by Class Rank, Winter and Spring Quarters, 1961-62 . . .	9
2. Correlation of Difficulties, Quizzes 5 and Quiz 1B	18
3. Test of the Significance of the Difference in Difficulty for Quizzes 5	19
4. Test of the Significance of the Difference in Difficulty for Quizzes 5B and 1B	19
5. Correlation of Discrimination Values, Quizzes 5 and Quiz 1b	20
6. Test of the Significance of the Difference in Total Test Discrimination for Quizzes 5	20
7. Test of the Significance of the Difference in Reliabilities for Quizzes 5	21
8. Difficulty and Discrimination Values of the Quiz 2 Items as Arranged for the Quiz 6 Selection Process	25
9. Test of the Significance of the Difference in Difficulty for Quizzes 6	27
10. Test of the Significance of the Difference in Reliabilities for Quizzes 6	28
11. Correlation of Difficulty and Discrimination Values for Quizzes 6	28
12. Test of the Significance of the Difference in Total Test Discrimination for Quizzes 6	29
13. Test of the Significance of the Difference in Difficulty for Quizzes 7	34
14. Test of the Significance of the Difference in Total Test Discrimination for Quizzes 7	34

TABLES (Continued)

	Page
15. Test of the Significance of the Difference in Reliabilities for Quizzes 7	35
16. Correlation of Difficulty and Discrimination Values for Quizzes 7	35
17. Correlations Between Obtained and Estimated Difficulties, Quizzes 7	37
18. Test of the Significance of the Difference Between Correlations Between Estimated and Obtained Difficulties, Quizzes 7	38
19. Mean Estimates of Difficulty and the Significance of the Difference Between the Mean Estimates of Difficulty for Quizzes 7	38
20. Test of the Significance of the Difference Between Estimates for the First Ten Items of Forms 7A and 7B	39
21. Difficulty and Discrimination Values of the Quiz 4 Items as Arranged for the Quiz 3 Selection Process	43
22. Test of the Significance of the Difference in Difficulty for Quizzes 8	47
23. Test of the Significance of the Difference in Total Test Discrimination for Quizzes 8	47
24. Test of the Significance of the Difference in Reliabilities for Quizzes 8	47
25. Correlation of Difficulty and Discrimination Values for Quizzes 8	48
26. Correlations of Discrimination Values for all the Readministrations	50
27. Test of the Significance of the Difference in Difficulty on Knowledge of Results Answer Boards for Quizzes 3	50
28. Test of the Significance of the Difference in Reliabilities on Knowledge of Results Answer Boards for Quizzes 8	51

TABLES (Continued)

Table	Page
29. Test of the Significance of the Difference in Test Discrimination on Knowledge of Results Answer Boards for Quizzes 8	51
30. Correlations of Difficulty and Discrimination Values on Knowledge of Results Answer Boards for Quizzes 8	52
31. Tests of Significance of Differences in Difficulty and Discrimination on Quizzes 8; IBM Answer Sheets vs. Knowledge of Results Answer Boards	53
32. Lund's Arrangement of Items	56

CHAPTER I

THE PROBLEM

In the construction of measurement instruments and tests, there are two attributes generally regarded as of the greatest importance: reliability and validity. Most measurement specialists agree that, once an adequate degree of reliability has been obtained, validity is the single most important attribute of a test. For example, Anastasi says "Undoubtedly the most important question that needs to be raised regarding any psychological test concerns its validity" (1, p. 29). Cronbach states that "The quality which most affects the value of the test . . . is its validity" (6, p. 103). He also says "The first important consideration in choosing a test is its validity" (6, p. 126). Thorndike and Hagen state the same thought by saying "The first and foremost question to be asked with respect to any testing procedure is: How valid is it?" (27, p. 160.)

A test or test item is valid to the degree that it correlates with an acceptable criterion measure. This criterion measure may be one of two major types: external or internal. External criteria are performances other than (although often quite similar to) the test performance. For

example, the validity of a college entrance examination might well be the degree to which students' scores on it predict (are correlated with) college point-hour ratio.

Internal criteria, on the other hand, are those which are based on the responses to the test instrument itself. The most common internal criterion is the total score on the test. The two major assumptions involved in the use of the total test score as the internal criterion are that (1) the test, and each of the items within it, are measuring one and only one ability or factor (homogeneity of function) and (2) that the item scores have been summated in a meaningful fashion to obtain the total score.

Most educational course examinations utilize internal criterion measures to indicate the effectiveness of individual items.

For most achievement tests . . . it is difficult or impossible to obtain (external) criterion measures . . . the context of an achievement test is often formulated by analyses of curriculums and textbooks and by the pooled judgment of recognized authorities in the field. Under these circumstances a well-constructed test may constitute the best available measure of the criterion; in a sense, the test itself defines the function it is to measure. Such tests may be described as self-defining (7, p. 286).

In this case, then, when the test is self-defining and an internal criterion is utilized, the item validity statistics used are those entitled "internal consistency discrimination measures." The objective of such measures is to indicate the degree to which an item is differentially

responded to by those subjects with high total test scores as against those with low total test scores. It is, therefore, measuring the degree to which the item discriminates between subjects with high and those with low criterion scores.

Measurement of performance is difficult in many areas because the behavior evaluated is influenced by many different variables. The way in which a person performs in a measurement situation has been found to be related to a large number of influences other than the direct questions being asked.

The influence of past experience on present expectation has been shown by Bruner and Postman (2). They mixed a few incongruous playing cards (such as a red four of spades) with a series of normal cards. The incongruous cards, presented with the aid of a tachistoscope, were identified as normal (the red four of spades as a red four of hearts or a black four of spades) at a highly significant level on the first occurrence.

Another experiment that shows the influence of past experience utilized two-dimensional cutouts in the shape of a banana, tangerine, tomato and other objects. All the objects were cut from the same color paper and presented to a series of subjects who were then asked to match the color of the presented object by adjusting a color wheel. The

subjects distorted the matching in the direction of the natural color of the objects (3).

Instructions can also influence the manner of response; Siipola interspersed a half-dozen "nonsense" words among a list of normal words. One group of subjects was instructed that the list would be concerned with animals and birds while another group was told the list would have to do with travel and transportation. Siipola found that significantly more responses to the nonsense words were in the instructed direction (26).

Another demonstration of the importance of instructions comes from a discussion by Murphy:

. . . Marks, using an ergograph, constructed work curves for young adult males. The distribution was not normal, but flattened out. The explanation appeared to be in the form of the instructions: 'Work as hard as you can and as fast as you can.' About two-thirds of these men apparently heard these instructions with the emphasis upon 'hard' and the other one-third with the emphasis upon 'fast.' Since it is impossible to give full attention to speed and effort at the same time, some functioned in a 'hard'-instruction, others in a 'fast'-instruction situation (21, p. 385).

These studies are supported by other literature showing that interest and personality profiles can be manipulated by instructions to respond in a given fashion (see, for example, 16 and 17).

In the field of interviewing, bias due to the interviewer, as well as to the phrasing of the questions asked of the interviewee, has long been recognized as one of

the major problems in obtaining valid information by this method.

. . . There is an abundance of experimental evidence to prove that bias may result, under certain conditions, regardless of anything the interviewer may do to eliminate it. In one study, 50 per cent of a sample of non-Jewish respondents told non-Jewish interviewers that they thought Jews had too much influence in the business world, whereas only 22 per cent of an equivalent sample voiced that opinion to Jewish interviewers. Similar experiments have shown that Negroes will frequently answer differently when interviewed by white people, and that working-class respondents are less likely to talk freely to middle-class interviewers. . . . in one study, it was demonstrated that Negroes spoke more freely with white interviewers in New York than they did in Memphis, Tenn. (13, p. 473).

The preceding studies, plus many more of the same type, show conclusively that many factors external to the test are influential in determining the responses of subjects to the test material.

Factors internal to the test are also influential. A number of texts (see 27, pp. 60-85) have a detailed discussion of correct and incorrect (confusing, ambiguous, etc.) ways of writing test items that influence the difficulty and discrimination of the items and the test.

Another internal factor that has received much verbal and less experimental attention is the ordering of the items by their difficulties. For years, it has been stated, without empirical evidence, that test items should be arranged in order of increasing difficulty. This is exemplified by Ross who states, without evidence, that "the items

in the test should be arranged in order of difficulty. It is especially important to have the easiest items at the beginning and the hardest ones at the end of the test. The exact order of the intervening items is less important" (24, p. 117). Interesting rationalizations are also given by Ruch (25), Greene, Jorgensen, and Gerberich (11), and Remmers and Gage (23), still without empirical evidence.

Lund, in a recent dissertation, found five common reasons for belief in the practice of ordering items in terms of increasing difficulty:

1. Difficult items early in the test disrupt the intellectual functioning of the subject.
2. Encountering difficult items early in testing is likely to result in an unwise use of time and to thereby to lower the performance score.
3. The less able subject is unable to cope with the items and will be discouraged.
4. Normal arrangement of items is necessary to encourage naive subjects, but that sophisticated subjects would not be disconcerted by other arrangements.
5. The overall morale or motivation level of the subjects would be adversely affected by encountering difficult items early in the testing (19, pp-7-8).

Empirical evidence on this topic is found in the work of Mensh, Capron and Lund.

Mensh found that "Under conditions of group testing, changed order of difficulty affected mean scores on the basic items of the experimental test. Irregular order of difficulty resulted in poorer performance by the subjects than was manifested when the basic items were presented in increasing order of difficulty" (20, p. 35).

Capron, using three orders of item difficulty arrangement, (I) easy-to-hard, (II) hard-to-easy, and (III) random, found a "general tendency for order I to be slightly easier than order II or order III; such differences . . . were statistically significant only once . . ." (4, p. 694).

Lund (19) used two forms of an intelligence test, one on which the items were ordered in terms of increasing difficulty and the other on which the 25 per cent most difficult were placed at the beginning and the remainder in order of increasing difficulty. He found that the latter form showed a decrement in total score.

All the available evidence, then, tends to support the view that deviations from an increasing item difficulty order causes scores on the total test to be lower in some cases. And the researchers who obtained these results indicated that the increasing order is, therefore, the preferred one. But there is no evidence, or even mention, in the literature on the relationship of difficulty order to the validity (discrimination) of the items or of the test. And, as pointed out above, validity is considered one of the most important properties of a test.

An important consideration should be the effect of item difficulty order on item and test validity. No judgment should be made on the relative value of the different

ways of ordering items until evidence in this area is available. The purpose of this study was to obtain such information on the relationship between item difficulty orders and test validity.

CHAPTER II

PROCEDURE AND RESULTS

General Remarks

The study covered a period of two quarters, the winter and spring quarters of the 1961-62 school year. The winter quarter was used to construct original items for the course Educational Psychology 407 at The Ohio State University and to find the difficulty and discrimination values of those items. In the spring quarter, the items were rearranged according to their difficulties, as obtained in the winter, to test a number of questions concerning the relationship between order of difficulty and total test difficulty and between order of difficulty and total test discrimination. The spring quarter test administrations were given at four separate times and are referred to below as Experiments I, II, III, and IV.

The Course and the Text

Educational Psychology 407 is required of all education majors at the University and has one prerequisite: one quarter of General Psychology. There are a large number of 407 sections and instructors, no one instructor having more than two sections. Each section meets once a day, Monday

through Friday, for forty-eight minutes. Considerable freedom is allowed the instructors in their handling of the course; there are only two practices basic to all sections: (1) lecturing on the material in the text is regarded as undesirable and (2) the same quizzes are given to all sections on the same days by their regular instructors at the normal class hour. The text used in the course is Psychology in Education by Pressey, Robinson and Horrocks (22).

Quizzes in the Course

Four quizzes were given each quarter in 407, each quiz covering a major area of the field covered by the course. The quizzes are intended to evaluate the students' knowledge of, and ability to use, the facts and principles in the text. The first quiz evaluated four chapters from the text and the other three quizzes tested three chapters each. The quizzes were given two weeks apart on the third, fifth, seventh and ninth Mondays of the quarter.

Subjects

Most of the Psychology 407 students in the winter and spring quarters were freshmen and sophomores. The numbers and percentages at each class rank for the seventeen sections in the winter quarter and for the fifteen sections in the spring quarter are given in Table 1. As can easily be observed, the two populations were very similar in this respect.

TABLE 1

NUMBER AND PERCENTAGE OF 407 STUDENTS BY CLASS RANK,
WINTER AND SPRING QUARTERS, 1961-62

	Winter		Spring	
	Number	Per Cent	Number	Per Cent
Freshmen	291	44	244	42
Sophomores	220	33	213	37
Juniors	109	16	88	15
Seniors	36	6	25	4
Others	4	1	10	2
Total	660	100	580	100

Preliminary Procedure (Winter Quarter)

Construction of Items

For each of the four quizzes, eighty original items were written by the investigator. They were then examined by at least one of the other 407 instructors; those items and alternatives that were considered confusing, unimportant or non-discriminating were either rewritten or replaced.

The items were all of multiple-choice form with four alternatives and in all cases only one of the alternatives was considered to be the best (correct).

For the first quiz, twenty items were written for each chapter covered; for the other three quizzes there were either twenty-six or twenty-seven items per chapter.

Construction of the Forms

For each of the four quizzes the eighty items were divided into two forms of forty items each; the odd-numbered items (in order of being written) were placed on one form, the even-numbered items on the other. The items, and the pattern of correct alternatives, were arranged randomly through the use of a random numbers table.

Instructions

The instructions for Quizzes 1 are reproduced on page 13. They were revised very slightly for the next three quizzes to include the information that no penalty would be made for guessing.

Administration

Each instructor was given a folder containing all of the necessary materials, including tests, answer sheets, instructors, and test-scoring pencils. The two test forms were arranged alternately so that people sitting next to each other would have different forms.

Results

The difficulty (per cent passing, uncorrected for chance) and the discrimination value (point-biserial correlation between success on the item and total score on the test) were computed for each of the items through the use of an IBM test-scoring machine with a graphic item count attachment.

QUIZ INSTRUCTIONS

To the teacher: (You have 45 tests, 45 answer sheets and 50 test scoring pencils. Do not pass anything out or begin the instructions until the bell has rung and the students are seated)

To the students (after gaining their attention) "Put all your books, notebooks, pens, pencils, purses, etc. on the floor or on the coatracks." (While they are doing so, pass out one pencil and one answer sheet per student.) "Put your name, my name and the class hour on your answer sheet." (Wait a few seconds)

To the students: "There are a number of forms of this test and there are different items on each form. For all forms there is only ONE BEST ANSWER to each question. There are 4 possible answers to each question; if you think alternative a is the correct choice, fill in the space in column 1 on your answer sheet for the number of that item. If you think b is the correct alternative, fill in the space in column 2, etc. You will never have occasion to use column 5 on your answer sheet."

"There is a blue number on the upper right-hand corner of each test; put this number on the upper-right-hand corner of your answer sheet. I will not answer any questions about the test or about the definitions of the words in it. The only questions I will answer will be in cases where the words are unreadable. Do not turn the tests over until I tell you to do so." (Pass out the tests, one per student, face down)

(After the students have the tests) "Use the test-scoring pencils and erase all changes completely. Do not make any marks on the test! You may leave as soon as you finish. Are there any questions?"

(As soon as questions have been answered) "Put the blue number from the upper right-hand corner of your test on the upper right-hand corner of your answer sheet and begin working."

(After the students have begun, count the number of students present and record below)

Number of tests 45
Number of students _____

Test left over _____

(Check to see that you have the correct number of tests and answer sheets left over. If not, look around the tables and see if you can find the extras.)

(As the students leave, check their answer sheets to make sure there is a number on it and that the number corresponds to the test they hand in)

(Please get all the pencils back; they are expensive)

No correction for chance was employed because, in power tests with no omits, such as the tests in this study, such a correction would depress all difficulties proportionally to their uncorrected difficulties and leave their rank order unchanged. Inasmuch as the "true" difficulties are unimportant to the design or interpretation of the results of the experiments in this study, the corrections would be unnecessary added statistics. It is true that Davis (8) found that correcting for chance appears to make the discrimination values somewhat more independent of their difficulty. But comparisons were not made here of items differing in difficulty as would be done when selecting a small number of items from a larger group for inclusion in a final form. In this study the interest was in comparing the discrimination of the same item or set of items arranged in different orders.

The point-biserial correlation was used as the measure of discrimination for the items because it utilizes all of the available information (which the short-cut methods that use information only from some of the extreme cases do not) and because the biserial correlation would require the assumption of a normal, continuous distribution underlying

the pass-fail dichotomy. This assumption was felt to be too tenuous for the slight advantage that the biserial correlation might offer.

The items, listed by quiz number and by item number within the quiz, with their obtained difficulty and discrimination values, are given in Appendix A.

The means and standard deviations of each of the eight forms are presented in Appendix B.

Experiment I

Rationale

The possibility of testing the effect of ordering items on one test form by the difficulty values obtained on a previous administration rests on the assumption that the difficulty values stay relatively constant, or that, at the least, their rank order remains highly consistent.

Previous investigators have found item difficulties to be quite reliable. Hilgard and Jandron (12) gave a final examination in General Psychology at two separate schools with fifty per cent of the items common to both tests. They obtained a correlation of .80 for the relative difficulties of these items at the two schools. Carter (5), interested in the influence of sample size on the reliability of difficulties, found an average correlation of .93 for four groups of twenty-five students each and .96 and .97 for two groups of fifty. Davis (8), correlating difficulty indices from

two independent samples of 370 subjects each, found a relationship of .98. And Gibbons (10) arranged the items on an algebra scholarship test randomly one year and in order of increasing difficulty the following year. With populations of approximately 425 students, he found the reliability of the difficulties to be .97.

It would appear, from these studies, that items on the spring quarter quizzes could be arranged by difficulty, using the difficulties obtained in the winter quarter as the basis for the arrangement. Rather than rely completely on historical data, however, it was decided to empirically check the reliability of the difficulty values, as well as the discrimination values, by administering one of the winter quarter quiz forms in the spring quarter in exactly the same order as given in the winter. Form 1B was selected because (1) it had the greater difficulty range and (2) Form 1A had an item that was a negative discriminator at the .05 level of significance. Hereafter, Form 1B as given in the spring quarter will be referred to as Form 5B.

Construction of Forms

The items on Form 1B (and therefore Form 5B) had been arranged randomly; in order to test a further question during the same quiz administration, two other forms were built and administered at the same time. The two, Forms 5A and 5C, were composed of the same forty items contained in Form 5B.

The items on Form 5A were ordered in terms of increasing difficulty, from the easiest item first to the hardest item at the end; Form 5C was the exact reverse of 5A, with the hardest item first and the easiest item last.

These orders were used to gain information concerning (1) the differences in test difficulty and (2) the differences in discrimination of the tests resulting from these three different orderings.

From the results of the studies presented earlier on the effect of ordering by difficulty, it would be expected that Form 5A would be the easiest and Form 5C the hardest of the three. Due to the lack of earlier work, no hypotheses were made concerning the direction of differences in discrimination.

Instructions

In order to make the comparisons between Form 1B and 5B as valid as possible, the instructions for Quizzes 5 were the same as those used for Quizzes 1.

Results

The results of Experiment I are presented in Tables 2 through 7. As was predicted, the correlation between the obtained difficulties for Form 1B and Form 5B was very high, .979. This extremely high correlation indicates that the items kept their relative difficulty position from one

quarter to the next. This means, therefore, that the arrangement of items by their difficulties on one administration can be done with the assurance that, on a following administration with a different sample, the items will retain much the same difficulty order.

The similarly high correlations between the obtained difficulties of the three Quiz 5 forms is evidence that, regardless of the ordering of the items by their difficulties, the relative difficulties are quite consistent.

TABLE 2
CORRELATION OF DIFFICULTIES,
QUIZZES 5 AND QUIZ 1B

Form				
Form	5A	5B	5C	1B
5A		.978	.976	
5B	.978		.976	.979
5C	.976	.976		

The insignificance of the differences between the means of the three forms of Quiz 5 indicates that none of the three orders is distinctly more difficult than the others. This finding was in contrast to expectations derived from past studies.

TABLE 3
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE IN
DIFFICULTY FOR QUIZZES 5

Statistic	Form			
	A	B	C	
Mean	21.177	21.038	20.903	$t_{5A-5C} = .654$
S.D.	3.966	3.873	4.109	$.60 > p > .50$
N	186	184	185	

The mean score on Form 5B was higher than the mean score on Form 1B (Table 4). This difference was significant between the .10 and .05 levels of probability and indicates, at this level, that the students in the spring quarter found the test slightly easier than did the students in the winter.

TABLE 4
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN DIFFICULTY FOR QUIZZES 5B and 1B

Statistic	Form		
	5B	1B	
Mean	21.038	20.403	$t_{5B-1B} = 1.684$
S.D.	3.873	4.068	$.10 > p > .05$
N	184	325	

With respect to discrimination, the results contained in Tables 5 and 6 indicate that (1) the correlation of discrimination values from form to form was not extremely high (.479 to .521), but all three correlations were signifi-

cantly greater than zero beyond the .01 level, and (2) there were no significant differences between the discrimination power of the three forms, as evaluated by a matched-groups t test.

TABLE 5
CORRELATION OF DISCRIMINATION VALUES,
QUIZZES 5 AND QUIZ 1B

Form	Form			
	5A	5B	5C	1B
5A		.521	.507	
5B	.521		.479	.545
5C	.507	.479		

TABLE 6
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE IN TOTAL TEST
DISCRIMINATION FOR QUIZZES 5

Form				Form			
	A	B			B	C	
Mean	.220	.232	$t = .625$	Mean	.218	.232	$t = .771$
d	.419		$p > .50$	d	.559		$p > .40$
d ²	.501			d ²	.521		

The reliabilities of the three forms, obtained by use of the Kuder-Richardson Formula 8 (15) were .578, .553, and .598 for Forms A, B, and C respectively. The largest difference, that between .598 and .553, was not significant, even at the .75 level of significance.

TABLE 7
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN RELIABILITIES FOR QUIZZES 5

	Form			
	A	B	C	
KR 8	.578	.553	.598	$z_{5C-5B} = .289$
z'	.659	.626	.693	$p > .75$
N	40	40	40	

Experiment II

Rationale

Experiment I involved the comparison of three differently-ordered forms: (A) an increasing form, (B) a random form, and (C) a decreasing form. Experiment II was designed to compare two slightly different forms: (A) a form with a series of easy items at the beginning and the remainder of the items arranged randomly, and (B) a form with an initial series of hard items and the remaining items ordered randomly.

If one of these two forms were found to be more valid or reliable than the other, it would have implications for

classroom teachers. These designs could be approximated with less computational effort than those requiring a complete ranking of difficulty. To construct a test corresponding to those in Experiment II, the teacher would only have to be able to identify the most difficult, or the easiest, items from the others. And past experimentation indicates that teachers can evaluate relative difficulty well enough to do this task adequately.

Tinkelman (28) reports on two studies in which groups of judges estimated the difficulty of a series of arithmetic achievement items. In one study, thirty members of the examining staff of a public personnel agency judged the difficulty of one hundred items. The correlations between judgments and obtained difficulties ranged from .23 to .77 with a median value of .53. In the second experiment, Tinkelman used nine graduate students in an advanced test construction course. Again evaluating relative difficulty, he found a median r of .55.

Lorge and Kruglov (13) have also investigated in this area. They used experienced schoolteachers of mathematics, all of whom were candidates for advanced degrees, as judges. These judges estimated the difficulties for all the items of a 45-item mathematics test. For two groups of seven judges each, they obtained correlations between the estimated and empirical ranks of the difficulty of the items. These correlations ranged from .515 to .793 with average r_s of .65 and .72.

Construction of the Forms

Two forms were constructed for Experiment II: Form 6A, in which the ten easiest items (of the eighty covering the material for the area) would be placed at the beginning of the quiz, and Form 6B, in which the ten hardest items would be placed at the beginning of the quiz.

Further requirements were included in these forms:

1. In order to make a more valid comparison between forms possible, both forms should be composed of the same forty items.

2. The items in the quiz should have difficulties representative of the complete range of possible difficulties.

3. To test the effect of the series of items of rather homogeneous difficulty on items immediately following the series, an item whose difficulty is approximately halfway between the average difficulty of the ten easy items and the average difficulty of the ten hard items should be placed in position eleven on both forms.

4. The quizzes should give approximately equal weight to each of the three chapters evaluated.

5. The most important principles in the area tested must be included in the quiz.

6. The better discrimination values should be included whenever a choice was possible.

The eighty items used in Quizzes 2 were ordered according to their difficulties. Table 8 shows this

arrangement with the item number, difficulty, discrimination value, and chapter for which the item was written. Difficulty is given in terms of the percentage passing the item.

The first step was to select the ten easiest and the ten hardest items. The ten items at the top and the ten at the bottom of Table 8 were originally selected. Only one change was made in these; item B29 was replaced by item B16 as a representative of the ten hardest items because of B29's negative (though insignificant) discrimination value.

The remainder of the items were divided into difficulty levels, each level including a range of .05. Two items were selected from each range in such a manner as to fulfill requirements 2,4,5, and 6 above.

The average difficulty of the ten selected easy items was .927 and the average difficulty of the ten selected hard items was .220. The difficulty value midway between these two values would be .5735. The previously selected item closest to this value was item B26 with a difficulty of .580. Item B26, then, was picked to be item eleven for both forms and fulfill requirement 3 above.

For Form A, the initial group of ten easy items and then the remaining group of twenty-nine items was randomly arranged through the use of a random numbers table.

For Form B, the initial group of ten hard items and then the remaining group of twenty-nine items were randomly arranged through the use of a random numbers table.

TABLE 8

DIFFICULTY AND DISCRIMINATION VALUES OF THE QUIZ 2 ITEMS
AS ARRANGED FOR THE QUIZ 6 SELECTION PROCESS

Item: Form and Number	Chap- ter	Diff- culty	Discrim- ination	Item: Form and Number	Chap- ter	Diffi- culty	Discrim- ination
*A 5	8	.075	.075	A 19	9	.400	.099
B 29	8	.141	-.047	*A 28	8	.406	.280
*B 19	7	.157	.215	*A 38	9	.409	.463
*B 27	8	.172	.055	A 27	9	.419	.392
*A 22	7	.203	.343	B 30	7	.420	.272
*A 14	8	.231	.130	A 6	7	.431	.280
*B 9	8	.235	.132	<u>B 39</u>	<u>7</u>	<u>.442</u>	<u>-.139</u>
*A 2	8	.259	.077	B 23	9	.455	.401
*B 1	9	.266	.224	*B 31	8	.461	.405
*B 18	9	.292	.347	A 24	9	.462	.294
A 13	8	.303	.113	*B 3	8	.464	.401
<u>*B 16</u>	<u>8</u>	<u>.307</u>	<u>.265</u>	B 8	9	.473	.395
A 23	8	.341	.094	A 25	9	.484	.202
B 17	7	.348	.243	A 21	8	.488	.217
*B 10	8	.354	.211	<u>A 40</u>	<u>7</u>	<u>.494</u>	<u>.317</u>
B 34	8	.354	.231	*B 21	7	.511	.202
A 33	7	.369	.196	*A 12	7	.528	.235
*B 28	9	.373	.166	<u>B 12</u>	<u>7</u>	<u>.536</u>	<u>.156</u>
<u>B 40</u>	<u>7</u>	<u>.398</u>	<u>.027</u>	*A 11	7	.553	.354

TABLE 8 (continued)

Item: Form and Number	Chap- ter	Diffi- culty	Discrim- ination	Item: Form and Number	Chap- ter	Diffi- culty	Discrim- ination
B 14	9	.577	.312	*B 6	9	.762	.334
*B 26	9	.580	.390	*A 32	7	.772	.274
B 25	9	.589	.286	A 39	8	.784	.198
<u>A 26</u>	<u>7</u>	<u>.597</u>	<u>.265</u>	<u>B 33</u>	<u>9</u>	<u>.787</u>	<u>.179</u>
A 8	9	.600	.280	*B 7	7	.812	.237
*A 30	7	.619	.391	*A 15	9	.819	.295
*A 3	9	.625	.414	<u>B 36</u>	<u>8</u>	<u>.843</u>	<u>.220</u>
<u>A 17</u>	<u>7</u>	<u>.647</u>	<u>.263</u>	B 5	8	.853	.231
A 16	8	.656	.187	B 4	7	.856	.150
*B 2	7	.665	.374	A 4	9	.866	.195
A 7	3	.666	.268	<u>A 9</u>	<u>8</u>	<u>.875</u>	<u>.167</u>
A 31	7	.666	.214	*B 13	9	.878	.175
B 11	3	.671	.348	*B 24	7	.878	.262
B 15	7	.677	.254	*A 37	7	.909	.113
*A 34	8	.688	.330	*B 22	7	.915	.112
<u>B 35</u>	<u>8</u>	<u>.696</u>	<u>.235</u>	*A 36	7	.928	.101
*A 35	7	.703	.403	*B 32	8	.931	.315
A 20	9	.725	.297	*A 29	9	.931	.122
<u>*B 20</u>	<u>9</u>	<u>.743</u>	<u>.370</u>	*B 37	9	.962	.122
B 38	9	.752	.147	*A 1	9	.966	.153
A 18	9	.759	.208	*A 10	8	.969	.140

*The item was selected for use in quizzes 6A and 6B.

Instructions

The instructions used in the administration of Quizzes 6 were the same as used for Quizzes 2, 3, and 4.

Results

The comparison of difficulties (Table 9) was consistent with the finding of Experiment I; the difference between the two means was very slight and could be expected to occur as often as 70 per cent of the time by chance alone.

TABLE 9
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN DIFFICULTY FOR QUIZZES 6

Statistic	Form		
	A	B	
Mean	24.036	23.933	$t = .258$
S.D.	4.934	4.372	$p > .70$
N	275	267	

The difference in reliabilities for the two forms was .03, a difference small enough to occur 45 per cent of the time by chance (see Table 10), a result that is also consistent with that found in Experiment I.

The correlation of difficulties, given in Table 11, was again much higher than the correlation of discrimination values, the two relationships being .967 and .702 respectively.

TABLE 10
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN RELIABILITIES FOR QUIZZES 6

	Form		
	A	B	
KR 8	.753	.674	$z = .690$
z'	.979	.819	$p > .45$
N	40	40	

TABLE 11
CORRELATION OF DIFFICULTY AND DISCRIMINATION
VALUES FOR QUIZZES 6

Difficulty:	$r(6A)(6B) = .967$
Difficulty:	$r(2)(6A) = .936$
Discrimination:	$r(6A)(6B) = .702$
Discrimination:	$r(2)(6A) = .315$

The one result of Experiment II that differed from the findings of Experiment I concerned the difference between the discrimination of the two forms. As seen in Table 12, Form A (with the easy items at the beginning) was significantly more discriminating than Form B at the .05 level of confidence. This result differs from that of Experiment I in which the decreasing form, 5C, was more discriminating (although far from significantly so) than the increasing form 5A. This distinct reversal was surprising to the investigator.

TABLE 12

TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN TOTAL TEST DISCRIMINATION FOR QUIZZES 6

	Form		
	A	B	
Mean	.283	.250	$t = 2.364$
d	1.317		$.05 > p > .02$
d^2	.346		

When Quizzes 6 were constructed, the same item was placed in position eleven on both forms to see if the two series of items of extreme difficulties had any noticeable effect on an item immediately following the series. It appears that there was none; on Form A the difficulty and discrimination were .524 and .234 respectively; on Form B they were .581 and .285.

Experiment II also made possible a comparison that could not be made with the data of Experiment I; the forty items used in Experiment II were taken from two separate quizzes of forty items each. This meant that the set of items determining the total test score, upon which the difficulty and discrimination indices were computed, were a partially different set than those used in the preliminary procedure. In an attempt to evaluate the effect of such a chance in the makeup of the tests, correlations were computed between the difficulties and the discrimination

values of the forty items on Quizzes 6A and the difficulties and the discrimination values of the same items on Quizzes 2. The correlation of difficulties was almost perfect, .986, and the correlation of discrimination values was also quite high, .315. Both of these were higher than the correlation of the two forms with each other, but the differences were small enough to be expected more than 20 per cent of the time by chance.

The high correlation for the difficulties indicates that the difficulty of an item is relatively unrelated to the other items in the same test.

The relationship of .315 obtained for the discrimination values was much higher than might be expected. It is significantly higher, at the .05 level, than the .545 obtained when the discrimination values from Quizzes 1B and 5B were correlated. It is also higher than might be expected from earlier studies. Carter (5), Gibbons (10), and Davis (8), using discrimination indices based on partial data (such as the upper and lower 27 per cent), found reliabilities for their indices of only .46, .58, and .65 respectively.

The size of the difference in correlations is even more surprising in view of the fact that the values related were item-test correlations; it would be expected that two sets of item-test correlations obtained on different groups of items and rearranged by difficulty, would be related to

a lesser degree than those obtained on the same group of items and not rearranged.

Experiment III

Rationale

The failure to find significant differences in Experiment I led the investigator to wonder, and hypothesize, why no significance had been found, either in the difficulty or the discrimination of the different orders. (Because of typing and clerical deadlines, the results of Experiment II were not yet available when Quizzes 7, used in Experiment III, were constructed.) One of the questions that arose was, Is it possible that students do not recognize the difficulty of the items? Don't they recognize hard items as more difficult than easy items?

In the discussion of rationale for Experiment II, evidence was given that experienced teachers can evaluate relative difficulty fairly well (page 21). In an effort to answer the questions posed above, Experiment II attempted to obtain (1) information concerning students' ability to estimate the difficulty of the items while taking the test and (2) a second comparison between forms with items in increasing as opposed to decreasing orders of difficulty.

Construction of the Forms

Two forms were again constructed for Quizzes 7; Form A, on which the items were arranged in increasing order of

difficulty, and Form B, on which the same forty items were arranged in descending order of difficulty.

The items used in Forms 7A and 7B were those used earlier on Form 3B in the winter quarter.

Instructions

The instructions for Quizzes 7 contained the same basic information as in the previous quiz instructions plus the request for the student to estimate the difficulty of the items. These instructions are reproduced on page 33.

Results

The results of Experiment III are divided in two sections: (1) the evaluation of the difficulty, reliability, and discrimination of the two Quiz 7 forms and (2) the results of the students' estimation of the difficulty of the items.

Comparison of Quizzes 7. As in Experiment I, the reliabilities, difficulties, and discrimination values of the two quiz forms for Experiment III were not significantly different. A difference as large as that obtained between the mean difficulties for the two forms could be expected to occur as frequently as 60 per cent of the time as a result of chance alone (Table 13), and a difference as large as that found between the discrimination values could be expected at least 20 per cent of the time by chance.

INSTRUCTIONS FOR QUIZZES 7

To the teacher (You have 45 tests, 45 answer sheets and 50 test-scoring pencils. Do not pass anything out or begin the instructions until the bell has rung and the students are seated)

To the students "Put all your books, notebooks, pens, pencils, purses, etc. on the coat racks or on the floor" (While they are doing this, pass out one pencil and one answer sheet per student) "Put your name, my name, and the class hour on your answer sheet."

To the students "There are a number of different forms of this test. For all forms there is only ONE BEST answer for each question. Do not mark more than one answer for any question. There are 4 possible answers to each question; do not use column 5 on your answer sheet. There is no correction for guessing; your score will simply be the number of questions that you answer correctly."

"Immediately after marking your answer to each question, indicate how difficult you think the question is; do this by marking a number from 1 to 5, using the scoring system on the blackboard,¹ in the space to the right of the column 5 on the answer sheet.

"Use the test-scoring pencils and erase all changes completely. Do not make any marks on the test booklet. You are to mark the blue number on the upper-right hand corner of the test on the upper-right corner of your answer sheet before you begin. Are there any questions?"

(As soon as questions have been answered) "Put the blue number from the upper-right hand corner of your test on the upper-right corner of your answer sheet and begin working."

Number of tests 45
 Number of students _____
 Tests left over _____

¹The scoring system given on the blackboard was
 1: very easy
 2: easy
 3: average
 4: hard
 5: very hard

(Table 14). It might also be noted that the decreasing form had the better discrimination, as it did in Experiment I.

TABLE 13
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN DIFFICULTY FOR QUIZZES 7

Statistic	A	B	
Mean	26.135	26.330	$t = .400$
S.D.	5.504	5.750	$p > .60$
N	267	267	

TABLE 14
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN TOTAL TEST DISCRIMINATION FOR QUIZZES 7

	Form		
	A	B	
Mean	.309	.326	$t = 1.223$
d	.878		$p > .20$
d^2	.332		

The reliabilities of the two forms, .778 and .805 for Form A and Form B respectively were not significantly different. A difference as large as that obtained would be expected more than 70 per cent of the time by chance alone (Table 15).

TABLE 15
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN RELIABILITIES FOR QUIZZES 7

	Form		
	A	B	
KR8	.778	.805	$z = .315$
z'	1.040	1.113	$p = .70$
N	40	40	

Also consistent with the two earlier experiments were the correlation of difficulties and discrimination values between forms, displayed in Table 16. The difficulties of the items of Forms 7A and 7B correlated .952 and the discrimination values correlated .611

TABLE 16
CORRELATION OF DIFFICULTY AND DISCRIMINATION
VALUES FOR QUIZZES 7

Difficulty:	$r(7A)(7B) = .952$
Difficulty:	$r(3B)(7A) = .962$
Discrimination:	$r(7A)(7B) = .611$
Discrimination:	$r(3B)(7A) = .467$

Correlations were also computed between the difficulties and discrimination values of the items of Quizzes 3B and 7A in an effort to evaluate the consistency of results when items were used on two different samples as well as

rearranged by their difficulties on the second sample. These results are also found in Table 16. The correlation of difficulties continued to stay high, .962, and the discrimination values were related .467. These values are quite similar to the correlations obtained when the items were readministered to a different sample in the same order (Experiment I, p. 18, Table 2, and p. 20, Table 5). In the earlier experiment, the values obtained were .979 and .545 respectively.

Students' estimation of difficulty. The analysis of the students' ability to estimate the difficulty of the test items was accomplished in the following manner: (1) The papers on each of the two Quiz 7 forms were divided into thirds by total score on the quiz; (2) the mean estimate of difficulty was obtained for each of the items for the upper and lower thirds on each of the forms; (3) correlations were computed between the obtained difficulties of the items and the estimated difficulties of the same items (the obtained difficulties being based on all the students who took that form); and (4) significance tests were computed for the difference between the difficulty estimates of the upper third and the lower third for each of the forms.

The obtained correlations, presented in Table 17, indicate that all four groups estimated the obtained difficulties significantly better than chance, the correlations

ranging from .461 to .701, with a correlation of .280 being sufficient for significance at the .01 level.

TABLE 17
CORRELATIONS BETWEEN OBTAINED AND
ESTIMATED DIFFICULTIES, QUIZZES 7

		Obtained	
Estimated By		7A	7B
7A	Upper Third	.643	
	Lower Third	.701	
7B	Upper Third		.595
	Lower Third		.461

As may be observed from Table 18, the correlations between the estimated and obtained difficulties for the two extreme thirds were not significantly different on either form. This would indicate that the best and the poorest students, on the basis of total score, evaluate relative difficulty equally well.

Reference to Table 19 reveals that for both forms students with total scores in the lower third thought the items were significantly more difficult than did the students in the upper third, the level of significance being .02 for Form B and .01 for Form A.

The answer to the questions posed at the beginning of the experiment appears to be that the students taking part in these studies can, regardless of their knowledge of

the material, estimate the relative order of the difficulties of the items while they answer those items, and that the students with greater knowledge of the material estimate the items to be easier than do those with less knowledge.

TABLE 18

TEST OF THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN
CORRELATIONS BETWEEN ESTIMATED AND OBTAINED
DIFFICULTIES, QUIZZES 7

Correlation Between				
	7A		7B	
	Upper Third and Obtained	Lower Third and Obtained	Upper Third and Obtained	Lower Third and Obtained
r	.643	.701	.595	.461
z'	.763	.869	.685	.499
N	40	40	40	40
z =	.457		.802	
p >	.60		.40	

TABLE 19

MEAN ESTIMATES OF DIFFICULTY AND THE SIGNIFICANCE OF THE
DIFFERENCE BETWEEN THE MEAN ESTIMATES OF DIFFICULTY
FOR QUIZZES 7

	7A		7B	
	Upper Third	Lower Third	Upper Third	Lower Third
Mean	3.20	3.32	2.96	3.06
t	3.055		2.658	
p	p < .01		.02 > p > .01	

At this point, one further question was investigated. It seemed possible that, although the students could estimate fairly accurately the relative difficulty of the items, they might not be able to tell at the beginning of a test whether the items they were answering at that time were harder or easier than average. If this were true, it would be expected that the estimation of difficulty for the first few items on the two forms would not differ significantly, but that the students would use these early items as a standard and estimate the difficulty of later items with reference to this standard. They could, in this manner, estimate the relative order of difficulties of all the items without being able to perceive whether the first few items were extremely hard or extremely easy.

As a test of this, the significance of the difference between the mean difficulty estimate for the first ten items of the two forms was determined. The result (see Table 20) was not significant at the .15 level of confidence.

TABLE 20
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
BETWEEN ESTIMATES FOR THE FIRST TEN
ITEMS OF FORMS 7A and 7B

	Form		
	A	B	
Mean	2.97	3.13	$t = 1.35$
S.D.	.36	.34	$p > .15$
N	10	10	

These results give support to the conclusion, therefore, that students cannot reliably tell whether they have started with the hardest or with the easiest items in a test until after completing quite a few of the items. And, if the students are unable to perceive the difficulty of the earliest items, differential perception of the task, and differential response due to this perception, could not be expected to occur at this point. This could account, at least in part, for the failure to find significance differences in the difficulty or discrimination of the differently ordered forms in Experiment I.

Experiment IV

Rationale

The search for a reason for the failure to find significant differences between orders in Experiment I, which led to the students' estimation of difficulties in Experiment III, also led to a further modification in Experiment IV. It was suggested to the investigator, by one of the 407 instructors, that if a student knew that he had failed a large percentage of the items at the beginning of a quiz it might cause him to respond differently than if he knew that he had passed a large percentage of the earliest items.

The decision to investigate this possibility led to the construction of an immediate-knowledge-of-results answer board and its use in four of the fifteen sections. This

form was used in sections meeting late in the day so that knowledge gained from its use could not be passed on to students in the other sections.

Construction of the Answer Board

The answer form constructed for use consisted of a 7x11-inch piece of eight-ply posterboard, a 9x11-inch sheet of carbon paper and an 8-1/2 x 11 inch dittoed answer sheet. The correct answer patterns were determined for the two forms of Quiz 8 and X's indented in the posterboard under the places where those answer squares would appear on the answer sheets. When the square on the answer sheet corresponding to the correct answer was filled in by rubbing lightly with a pencil, the X would appear in the square. Filling in one of the incorrect answer squares resulted in a solidly filled square with no mark appearing in it. Red crayon pencils were furnished with the boards because the X's showed up more clearly than when black lead pencils were used.

The answer sheet was stapled to the board with the carbon paper between the two and facing such that all marks would be revealed on the back of the answer sheet. The carbon paper was placed in this fashion in order to detect any searching, for correct answers, with fingernails or other objects not leaving a mark on the front of the answer sheet.

Construction of Forms

The only difference between Quizzes 5 and 7, used in Experiments I and III, and Quizzes 6 (on which the significant difference in discrimination was found) was that the ten hardest and ten easiest items on Quizzes 6 were slightly more extreme and homogeneous with respect to difficulty than were the same groups of items on Quizzes 5 and 7. This is because they were selected with that goal in mind from both forms of Quizzes 2. In order to make Quizzes 8 more like Quizzes 6, the same criteria were used for the inclusion of items in Quizzes 8 that were previously employed in constructing Quizzes 6 (see page 23) with one exception; requirement 3 (placing the same item in position eleven on both forms) was omitted.

After the items were selected, two forms were constructed: Form 8A, on which the items were arranged in increasing order of difficulty, and Form 8B, on which the same forty items were arranged in decreasing order of difficulty.

The items from which the selection was made, and the selected items (marked with an asterisk) are given in Table 21.

Instructions

For the sections not using the knowledge of results answer boards, the instructions were the same as for Quizzes 2,3,4, and 6. The instructions used with the answer boards are given on page 45.

TABLE 21

DIFFICULTY AND DISCRIMINATION VALUES OF THE QUIZ 4 ITEMS
AS ARRANGED FOR THE QUIZ 8 SELECTION PROCESS

Item: Form and Number	Chap- ter	Diffi- culty	Discrim- ination	Item: Form and Number	Chap- ter	Diffi- culty	Discrim- ination
*B 31	15	.951	.217	*B 8	3	.777	.316
*B 30	13	.905	.304	B 17	15	.777	.280
*B 19	15	.879	.336	B 3	15	.774	.178
*A 26	3	.873	.305	*A 6	15	.773	.472
*A 35	13	.870	.305	<u>A 40</u>	<u>15</u>	<u>.759</u>	<u>.331</u>
*B 29	3	.869	.062	*A 37	13	.746	.360
*A 39	15	.863	.320	*B 4	3	.728	.316
*A 29	3	.860	.272	A 11	3	.726	.316
*B 1	13	.859	.148	B 23	13	.721	.237
<u>*A 31</u>	<u>13</u>	<u>.856</u>	<u>.342</u>	<u>B 26</u>	<u>3</u>	<u>.715</u>	<u>.162</u>
A 14	15	.819	.295	A 10	15	.699	.357
*B 14	3	.813	.279	A 20	3	.686	.382
B 11	13	.813	.222	*A 30	15	.686	.308
A 22	15	.806	.266	B 9	3	.685	.213
<u>*A 33</u>	<u>3</u>	<u>.806</u>	<u>.201</u>	B 39	15	.672	.280
A 23	3	.799	.412	*B 13	13	.669	.444
B 33	3	.784	.128	B 40	13	.666	.370
A 38	3	.783	.292	B 22	3	.656	.149
B 5	15	.777	.203	<u>*A 13</u>	<u>13</u>	<u>.652</u>	<u>.454</u>

TABLE 21 (Continued)

Item: Form and Number	Chap- ter	Diffi- culty	Discrim- ination	Item: Form and Number	Chap- ter	Diffi- culty	Discrim- ination
B 7	13	.649	.358	A 8	13	.515	.191
A 2	15	.642	.206	*B 15	13	.511	.294
*B 24	3	.639	.417	B 16	15	.505	.493
B 32	15	.639	.268	<u>A 16</u>	<u>13</u>	<u>.502</u>	<u>.302</u>
*B 2	15	.610	.249	B 37	3	.495	.239
<u>*A 17</u>	<u>15</u>	<u>.595</u>	<u>.393</u>	*A 1	13	.482	.306
A 32	13	.595	.371	*B 10	3	.449	.225
*A 15	3	.595	.092	A 5	3	.435	.104
A 34	3	.592	.245	B 23	3	.416	.302
*A 4	15	.582	.465	*B 6	15	.403	.305
A 13	13	.575	.333	B 25	15	.400	.280
A 19	3	.575	.230	*B 18	15	.387	.312
<u>A 21</u>	<u>15</u>	<u>.559</u>	<u>.335</u>	*A 7	13	.381	.422
B 35	13	.544	.215	*A 25	3	.361	.248
B 36	13	.541	.379	*A 12	13	.314	.336
B 21	13	.541	.232	*A 23	15	.241	.357
A 24	13	.538	.240	*B 34	13	.197	.242
*A 3	15	.528	.452	*A 27	13	.194	.258
*B 33	15	.521	.404	*A 36	13	.151	.051
B 27	13	.521	.236	*B 12	3	.144	.170
B 20	13	.515	.282	*A 9	3	.094	.112

*The item was selected for use in quizzes 8.

D is the best answer. If you rub your pencil over the fourth answer square from the left you should see an X. Try it. Any questions?

Try the next two. (pause so they can) You should have found an X in the c square for example B and in the b square for example C. Any questions?

Write your name, my name and the class hour on the edge of the form and begin.

Results

The analysis of the results of Experiment IV was handled in three parts: (1) the analysis of the responses of the students who answered the test on IBM answer sheets, (2) the analysis of the responses of the students who used knowledge of results answer boards, and (3) a comparison of the results of the two above analyses.

Responses on IBM answer sheets. These results were again consistent with those of Experiments I and III. The difference in the mean difficulty of the two forms could be expected to occur at least 30 per cent of the time due to chance (Table 22); the difference in test discrimination could be expected to occur more than 70 per cent of the time on the same basis (Table 23); and the difference in the reliabilities was not significant even at the .75 level (Table 24).

The significance of the difference in discrimination found in Experiment II did not, therefore, appear in this analysis. Form B was slightly more discriminating in this analysis, a result consistent with Experiments I and III,

and the opposite of what would be expected on the basis of the results from Experiment II.

TABLE 22
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN DIFFICULTY FOR QUIZZES 8

Statistic	Form		
	A	B	
Mean	23.698	24.168	$t = .964$
S.D.	4.803	4.916	$p > .30$
N	192	191	

TABLE 23
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN TOTAL TEST DISCRIMINATION FOR QUIZZES 8

	Form		
	A	B	
Mean	.284	.289	$t = .299$
d		.310	$p > .70$
d^2		.491	

TABLE 24
TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN RELIABILITIES FOR QUIZZES 8

	Form		
	A	B	
KR 8	.736	.747	$z = .103$
z'	.942	.966	$p > .90$
N	40	40	

The correlations of difficulties and discrimination values between the two forms, contained in Table 25, were also similar to the values obtained in the earlier experiments, being .979 and .522 respectively.

TABLE 25
CORRELATION OF DIFFICULTY AND DISCRIMINATION
VALUES FOR QUIZZES 8

Difficulty:	$r(8A)(8B) = .979$
Difficulty:	$r(4)(8A) = .974$
Discrimination:	$r(8A)(8B) = .522$
Discrimination:	$r(4)(8A) = .696$

The same situation existed in this experiment, with respect to the selection of items for inclusion, as it did in Experiment II; that is, selected items from two separate forms of a previous administration were used in one form in the Quiz 8 forms. Therefore, as was done in Experiment II, correlations were obtained between the difficulties and between the discrimination values of the items as obtained on Quizzes 4 and on Form 8A. The correlations were .974 for the difficulties and .696 for the discrimination values (see Table 25).

In Experiment II, an unexpectedly high correlation of .315 was obtained between the discrimination values of Form 6A and the values of the same items when used in Quizzes 2. The .696 found in this experiment for the same comparison,

using Form 8A and Quizzes 4 was again higher than the correlation of Forms 3B and 7A (.467) in which, as in Experiment I for the correlation of 1B and 5B, the same forty items were used on both forms. The difference between .696 and .467, however, is not significant at the .10 level of confidence, nor is the difference between .696 and .545, the relationship obtained earlier between Forms 1B and 5B.

In order to investigate this result further, all the situations in which the same items were readministered (in the same or a different order, and within the context of the same forty items or in a partially different set) were examined, and correlations between the sets of discrimination values were computed. In all, there were nine such situations and the results are presented in Table 26. Examination of this data shows that the correlations obtained earlier were the extreme ones and that the full set of correlations comprise a distribution that could be expected to arise by chance.

Knowledge of results answer boards. The difference between the difficulty of Form 8A and Form 8B when knowledge of results was present was not significant at the .05 level (Table 27), and the difference between the reliabilities of the two forms could be expected to occur at least 15 per cent of the time by chance (Table 28).

The difference in discrimination for the two forms (Table 29) was fairly substantial, however. It was large

TABLE 26

CORRELATIONS OF DISCRIMINATION VALUES
FOR ALL THE READMINISTRATIONS

Forms Correlated		Situation		Correlations Obtained
Preliminary Procedure	Experimental Procedure	Context	Arrangement	
1B	5B	same	same	.545
1B	5A	same	different	.638
1B	5C	same	different	.585
3B	7A	same	different	.467
3B	7B	same	different	.772
2	6A	different	different	.815
2	6B	different	different	.653
4	8A	different	different	.696
4	8B	different	different	.624

TABLE 27

TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN DIFFICULTY ON KNOWLEDGE OF RESULTS
ANSWER BOARDS FOR QUIZZES 8

	Form		
	A	B	
Mean	24.508	26.150	$t = 1.703$
S.D.	5.831	4.533	$p > .05$
N	59	60	

TABLE 28

TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN RELIABILITIES ON KNOWLEDGE OF RESULTS
ANSWER BOARDS FOR QUIZZES 8

	Form		
	A	B	
KR 8	.823	.727	$z = 1.047$
z'	1.165	.922	$p > .25$
N	40	40	

TABLE 29

TEST OF THE SIGNIFICANCE OF THE DIFFERENCE
IN TEST DISCRIMINATION ON KNOWLEDGE OF
RESULTS ANSWER BOARDS FOR QUIZZES 8

	Form		
	A	B	
Mean	.342	.274	$t = 2.273$
d		2.741	$.05 > p > .02$
d^2		1.599	

enough to be significant at the .05 level of confidence, with Form A being the more discriminating of the two. This finding agrees with that of Experiment II in finding the increasing form to be of significantly greater discriminating power.

The correlation of difficulties between the two forms remained high, .950, but the relationship of the

discrimination values dropped to .045, a relationship that could be expected to occur more than 10 per cent of the time by chance (see Table 30). This could be partially due to the use of a correlation coefficient as a measure of discrimination with such a small number of subjects (sixty on one form and fifty-nine on the other); correlations based on such small groups would not be expected to be exceptionally stable.

TABLE 30

CORRELATIONS OF DIFFICULTY AND DISCRIMINATION VALUES
ON KNOWLEDGE OF RESULTS ANSWER BOARDS FOR QUIZZES 8

Difficulty: $r(8A \text{ Boards})(8B \text{ Boards}) = .950$

Discrimination: $r(8A \text{ Boards})(8B \text{ Boards}) = .049$

Comparison of the results of the two answer modes.

The results of this analysis are contained in Table 31.

The difference between the difficulty of Form A when using the answer boards and when using the IBM answer sheets was small enough to be accounted for by chance more than 30 per cent of the time, but the difference between the two answer modes for Form B was quite large. The mean score obtained on the answer boards was enough greater to be significant at the .01 level of confidence.

The differences in test reliability between the answer modes were not significant at the .05 level; for Form A, the obtained difference could be expected more than

TABLE 31

TESTS OF SIGNIFICANCE OF DIFFERENCES IN DIFFICULTY AND
DISCRIMINATION ON QUIZZES 8; IBM ANSWER SHEETS VS.
KNOWLEDGE OF RESULTS ANSWER BOARDS

Difficulty						
	Form A			Form B		
	IBM Sheets	Answer Boards	Results	IBM Sheets	Answer Boards	Results
Mean	23.698	24.508		24.168	26.150	
S.D.	4.803	5.881		4.916	4.538	
N	192	59		191	60	
			$t = .964$			$t = 2.893$
			$p > .30$			$p < .01$
Reliability						
KR 8	.736	.823		.747	.727	
z^1	.942	1.166		.966	.922	
N	192	59		191	60	
			$z = .966$			$z = .190$
			$p > .30$			$p > .80$
Discrimination						
Mean	.284	.342	$t = 1.871$	.289	.274	$t = .714$
d		2.347	$p > .05$		.606	$p > .40$
d^2		1.640			.715	

10 per cent of the time by chance and the obtained difference for Form B could be expected three-fourths of the time by chance.

The most interesting result of the analysis concerns the discrimination of the forms; for Form 8A, the use of the knowledge of results answer boards produced a more discriminating test between the .10 and .05 levels; for Form B, however, the obtained difference was in favor of the answer sheets and was large enough to be expected by chance more than 40 per cent of the time.

CHAPTER III

SUMMARY AND INTERPRETATION OF RESULTS

Difficulty

Summary. For the six comparisons made in this study between the difficulties of differently ordered forms, as answered on IBM answer sheets, no difference was significant beyond the .30 level. For four such comparisons of increasing and decreasing difficulty forms, each order was easier than the other order twice. And both were easier than the random form in the one comparison made of these conditions.

When knowledge of results was present, the difference, in favor of the decreasing form, was not significant at the .05 level. In this case, students obtained higher scores on the answer boards than on the answer sheets for both forms, the difference being significant only for the decreasing order.

This last result is regarded as of doubtful validity, however, because the answer boards used in this study were new and untested. It is quite possible that methods unknown to the investigator were used to obtain correct answers.

Interpretation. The failure to find any significant differences in difficulty between different forms in this study is in direct contrast to the results obtained by

Lund (19). But there were procedural differences in the two studies that lead the investigator to accept, with reservations, Lund's theoretical conclusion.

Lund constructed two forms, one in which the items were in an increasing order of difficulty, and a second in which the hardest 25 per cent, ordered in terms of increasing difficulty, were placed at the beginning and the other 75 per cent were also arranged in an increasing difficulty order. Both forms contained the same 72 five-alternative multiple-choice items. His analysis of the results involved a comparison of the difficulty of the four separate quarters of the two forms, matching quarters with respect to the items included. He therefore compared the first quarter on the increasing form with the second quarters of the other form, etc. (see Table 32).

TABLE 32
LUND'S ARRANGEMENT OF ITEMS

Form	17-item quarters			
A-Increasing Order	1	2	3	4
B-25% Hardest in First Qtr.	4	1	2	3

Lund found no significant differences in his comparisons of quarters 1, 2, or 4, but he did find a significant difference between quarters 3, with scores being significantly higher on the increasing form.

Lund's interpretation of his results was that the hard items at the beginning of Form B had caused the students answering the form to use excessive energy in their attempt to overcome the disruptive effect of the early placement of the extremely hard items and that

The striking difference in performance on the third quarter of Form A as compared with the last quarter of Form B indicates that any energy mobilization that had compensated for the disruptive effect of encountering the hard items first had now been dissipated. Thus the motivational level was now inadequate to sustain effort at the level necessary to handle the difficult items in this section (19, p. 34).

This investigator feels that the following differences in the two studies could account for the discrepancy in results without invalidating Lund's conclusion:

1. Lund administered 82 items; this study used only 40. The fewer number of items used in this study might not have been enough to allow the disruptive effect of the hard initial items to have any significant effect on total score.

2. Lund's students were taking the test for counseling purposes; the students in the present study were taking tests that partially determined their grade in the course. The motivation of the students in this study could therefore be expected to be at a higher initial level, and consequently more resistant to the disruptive effect, than Lund's subjects.

3. Lund's Form B went from Hard to Easy and back to semi-Hard items, while the comparable forms in the present study went from Hard to Easy without any return to harder

items. The subjects in Lund's study who had been subjected to the disruptive effect of the hardest items early in the test therefore returned to hard items again at the end of the test, where the effects of disruption would be most noticeable. In this study, the subjects who encountered the most difficult items first did not return to further hard items after a series of intervening easy ones. Instead, they found easier and easier items as they continued, letting them expend less and less energy as they proceeded through the test.

These three points lead this investigator to conclude that Lund's conclusion concerning the cause for the difference in response to differently-ordered forms might operate in situations where one or more of the three above conditions existing in his study occur, but not under the conditions existing in this study.

Insofar as most classroom test situations are closer to the situation utilized in this study, it would seem that Lund's conclusion--that test items should be arranged in order of increasing difficulty--may be valid for long tests used in large-scale testing programs, but that order will make little difference in classroom tests constructed by the teacher for evaluation of her course objectives.

Reliability

Summary. In no case was the difference in reliability of the differently ordered forms significant. The

largest obtained difference could have occurred by chance more than 25 per cent of the time. The decreasing forms were more reliable than the increasing forms three times and the increasing forms were more reliable twice.

The use of the knowledge of results boards did not affect reliability, either. For the increasing form of Quizzes 8, the boards were more reliable than the answer sheets, but for the decreasing form the boards were less reliable, with neither of these comparisons being significant beyond the .30 level.

Interpretation. It would appear that there is no advantage gained by spending the time and effort to order items by difficulty if the effect on test reliability is used as the criterion.

Discrimination

Summary. For responses obtained on IBM answer sheets, which is comparable to the methods usually used in obtaining answers to classroom tests, the decreasing form was more discriminating on three occasions, but never beyond the .20 level. On the fourth comparison, the form with the easy items at the beginning was significantly more discriminating, at the .05 level, than the form on which the initial items were the hardest.

When knowledge-of-results answer boards were used, the increasing form was more discriminating than the

decreasing form, again between the .05 and .02 levels. However, the difference in discrimination between the answer boards and the IBM sheets was not significant beyond the .05 level for either order.

Interpretation. The inconsistency of results in this case is somewhat disturbing. Further experiments, both replicative and investigatory, need to be done in an effort to resolve the discrepancies obtained here.

The only cases in which the differences in discrimination were great enough to be significantly different occurred when forms beginning with groups of items with very extreme difficulties were compared. However, the normal classroom test does not contain as many extremely difficult items as were included in these forms, and tests constructed by testing agencies for large scale use are generally given with a time limit and are, therefore, not comparable to the tests in this study. For this reason, the results of the analysis of discrimination lead the investigator to the conclusion that, for classroom evaluation, no difference in test discrimination would appear to result from the way in which the teacher ordered her items as long as the students had adequate time to consider all the items.

Correlation of Difficulty Values

Summary. The correlation of difficulty values from form to form was quite high throughout the study. The range of values was from .950 to .986 for the eleven cases evaluated.

Interpretation. This extremely small range of very high values indicates convincingly that the difficulty of an item is a very reliable statistic regardless of the ordering of the items by difficulty or the specific content of the other items in the test (as long as all the items in the test are concerned with the same general topic).

Correlations of Discrimination Values

Summary. The range of correlations indicating the relationship between the discrimination values on the different forms was much greater than the corresponding range of difficulty values and the correlations were consistently much lower. This is consistent with earlier results found on the reliability of difficulty and discrimination indices (see the earlier discussions on pp. 15-16 and p. 30).

Interpretation. These low correlations indicate that frequently discrimination indices obtained on one administration of a test will vary considerably on a following administration. The relationships were consistently greater than chance, but a large amount of variance remains unexplained.

Student Estimates of Difficulty

Summary. Students in both the upper and lower thirds of the distribution of total scores on both the increasing and decreasing forms were able to evaluate relative difficulty significantly better than chance at the .01 level. However, for the first ten items, the difficulty estimation of the students who took the form with the easy items at the beginning was not significantly different than for the students who answered the form beginning with the most difficult items.

Interpretation. The students' inability to recognize the difficulty of the earlier items on a test indicates the futility of putting a series of items of homogeneous difficulty at the beginning of a test if the goal of such placement is to influence their motivation through their perception of the difficulty of the early items.

Suggestions for Future Research

The experiments carried out in this study were done with undergraduate college students in a required educational psychology course at a large state university. Multiple-choice items were used on all test forms, and all quizzes were administered by the regular instructors in a normal-length class period. The generalization of results is limited by these factors and the possibility exists that different results might be found if further research were done with different populations, materials, item forms and

test lengths. Some further studies that the investigator feels might be useful would involve--

1. Similar investigations on students in the elementary, junior high and secondary schools where the range of ability is greater and test-taking experience is less.

2. Similar investigations involving the use of completion or short answer item format. A student with only partial knowledge of the information required to successfully answer a question could believe that he had, either directly or through the elimination of the other alternatives, obtained the best answer on a multiple-choice item but realize that he had not produced the best answer for the completion or short answer item covering the same principle or facts. This difference in perception on the part of the students could cause a difference in the test statistics for these two item types.

3. Similar investigations in elective and advanced courses. Students are generally more interested in the material and less interested in "just getting by" in such courses. This could result in a difference in attitude and, therefore, response of the students to the tests in the two situations.

APPENDIX A

DIFFICULTY AND DISCRIMINATION VALUES FOR THE ITEMS IN THE QUIZZES USED IN THE STUDY

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FOR QUIZ 1, FORM A

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.498	.247	21	.890	.249
2	.790	.201	22	.850	.154
3	.774	.247	23	.690	.294
4	.354	.139	24	.480	.341
5	.448	.312	25	.781	.134
6	.643	.211	26	.881	.050
7	.213	.130	27	.806	.239
8	.922	.160	28	.288	.225
9	.950	.069	29	.668	.227
10	.799	.245	30	.533	.347
11	.539	.411	31	.665	.072
12	.661	-.086	32	.727	.435
13	.423	.205	33	.574	.238
14	.423	.222	34	.404	.312
15	.611	.321	35	.458	.225
16	.396	.164	36	.248	.334
17	.113	.102	37	.241	.270
18	.260	.142	38	.370	.271
19	.553	.244	39	.837	.235
20	.498	.250	40	.705	.323

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 1, FORM B

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.545	.216	21	.471	.173
2	.354	.215	22	.477	.332
3	.742	.236	23	.077	.023
4	.702	.238	24	.923	.091
5	.135	.204	25	.668	.099
6	.194	.194	26	.449	.084
7	.846	.257	27	.889	.315
8	.729	.297	28	.929	.163
9	.305	.038	29	.652	.214
10	.517	.315	30	.169	.206
11	.308	.342	31	.717	.197
12	.360	.337	32	.572	.335
13	.615	.163	33	.135	.164
14	.772	.279	34	.246	.326
15	.933	.177	35	.465	.310
16	.317	.096	36	.523	.166
17	.560	.283	37	.597	.424
18	.163	.155	38	.163	.140
19	.385	.301	39	.412	.281
20	.634	.409	40	.643	.326

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FOR QUIZ 2, FORM A

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.966	.153	21	.488	.217
2	.259	.077	22	.203	.343
3	.625	.414	23	.341	.094
4	.866	.195	24	.462	.294
5	.075	.075	25	.484	.202
6	.431	.280	26	.597	.265
7	.666	.268	27	.419	.392
8	.600	.280	28	.406	.280
9	.875	.167	29	.931	.122
10	.969	.140	30	.619	.391
11	.553	.354	31	.666	.214
12	.528	.233	32	.772	.274
13	.303	.118	33	.369	.196
14	.231	.130	34	.688	.330
15	.819	.295	35	.703	.403
16	.656	.137	36	.928	.101
17	.647	.263	37	.909	.113
18	.759	.208	38	.409	.463
19	.400	.099	39	.784	.198
20	.725	.297	40	.494	.317

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FOR QUIZ 2, FORM B

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.266	.224	21	.511	.202
2	.665	.374	22	.915	.112
3	.464	.401	23	.455	.401
4	.856	.150	24	.873	.262
5	.853	.231	25	.589	.236
6	.762	.334	26	.580	.390
7	.812	.237	27	.172	.055
8	.473	.395	28	.373	.166
9	.235	.132	29	.141	-.047
10	.354	.211	30	.420	.272
11	.671	.348	31	.461	.405
12	.536	.156	32	.931	.315
13	.878	.175	33	.787	.179
14	.577	.312	34	.354	.231
15	.677	.254	35	.696	.285
16	.307	.265	36	.843	.220
17	.343	.243	37	.962	.122
18	.292	.347	38	.752	.147
19	.157	.215	39	.442	-.139
20	.743	.370	40	.398	.027

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 3, FORM A

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.410	.285	21	.708	.096
2	.429	.363	22	.750	.250
3	.542	.389	23	.910	.110
4	.337	.184	24	.766	.342
5	.567	.275	25	.240	.271
6	.641	.195	26	.484	.360
7	.843	.140	27	.654	.249
8	.859	.198	28	.936	.147
9	.740	.064	29	.590	.058
10	.917	.296	30	.628	.039
11	.897	.141	31	.558	.143
12	.635	.422	32	.676	.347
13	.804	.236	33	.433	.431
14	.340	.199	34	.683	.238
15	.359	.202	35	.157	.149
16	.330	.353	36	.500	.363
17	.965	.292	37	.872	.249
18	.587	.367	38	.436	.358
19	.920	.166	39	.577	.379
20	.946	.212	40	.561	.319

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 3, FORM B

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.406	.251	21	.696	.420
2	.112	.214	22	.307	.231
3	.812	.240	23	.658	.293
4	.626	.185	24	.639	.343
5	.735	.298	25	.716	.359
6	.837	.284	26	.821	.321
7	.677	.311	27	.789	.290
8	.591	.189	28	.524	.217
9	.770	.339	29	.473	.364
10	.428	.278	30	.792	.352
11	.898	.267	31	.502	.339
12	.575	.293	32	.895	.384
13	.904	.130	33	.955	.230
14	.805	.355	34	.853	.164
15	.604	.435	35	.684	.290
16	.498	.283	36	.636	.388
17	.533	.144	37	.262	.096
18	.359	.230	38	.869	.310
19	.665	.503	39	.343	.228
20	.754	.366	40	.583	.354

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 4, FORM A

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.482	.306	21	.559	.335
2	.642	.206	22	.806	.266
3	.523	.452	23	.241	.357
4	.582	.465	24	.538	.240
5	.435	.104	25	.361	.248
6	.773	.472	26	.873	.305
7	.381	.422	27	.194	.258
8	.515	.191	28	.799	.412
9	.094	.112	29	.860	.272
10	.699	.357	30	.686	.308
11	.726	.316	31	.856	.342
12	.314	.336	32	.595	.371
13	.575	.333	33	.806	.201
14	.819	.295	34	.592	.245
15	.595	.092	35	.370	.305
16	.502	.302	36	.151	.051
17	.595	.393	37	.746	.360
18	.652	.454	38	.783	.292
19	.575	.230	39	.363	.320
20	.636	.382	40	.759	.331

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 4, FORM B

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.859	.148	21	.541	.232
2	.610	.249	22	.656	.149
3	.774	.178	23	.416	.302
4	.723	.316	24	.639	.417
5	.777	.208	25	.400	.280
6	.403	.305	26	.715	.162
7	.649	.358	27	.521	.236
8	.777	.316	28	.721	.237
9	.635	.213	29	.869	.062
10	.449	.225	30	.905	.304
11	.813	.222	31	.951	.217
12	.144	.170	32	.639	.263
13	.669	.444	33	.734	.128
14	.313	.279	34	.197	.242
15	.511	.294	35	.544	.215
16	.505	.493	36	.541	.379
17	.777	.230	37	.495	.239
18	.387	.312	38	.521	.404
19	.379	.336	39	.672	.280
20	.515	.282	40	.666	.370

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 5, FORM A

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.957	.103	21	.511	.337
2	.946	-.043	22	.522	.301
3	.898	.145	23	.522	.222
4	.892	.274	24	.511	.432
5	.376	.132	25	.414	.040
6	.801	.101	26	.489	.328
7	.790	.276	27	.360	.141
8	.747	.266	28	.452	.381
9	.667	.210	29	.306	.214
10	.726	.223	30	.312	.154
11	.651	.135	31	.344	.259
12	.796	.239	32	.387	.113
13	.651	.391	33	.247	.229
14	.715	.257	34	.103	.046
15	.629	.290	35	.210	.207
16	.640	.334	36	.194	.112
17	.554	.329	37	.172	.260
18	.613	.334	38	.113	.125
19	.457	.239	39	.138	.159
20	.656	.313	40	.151	.091

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 5, FORM B

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.560	.227	21	.473	.244
2	.283	.293	22	.484	.339
3	.793	.119	23	.136	.148
4	.766	.221	24	.891	.170
5	.103	.181	25	.679	.268
6	.196	.133	26	.348	.011
7	.342	.193	27	.897	.248
8	.788	.359	28	.924	.210
9	.342	.043	29	.674	.250
10	.478	.240	30	.158	.081
11	.364	.378	31	.728	.217
12	.462	.135	32	.565	.159
13	.647	.383	33	.158	.015
14	.728	.142	34	.201	.300
15	.957	.174	35	.603	.519
16	.386	.133	36	.636	.153
17	.598	.294	37	.701	.276
18	.250	.205	38	.114	.089
19	.380	.290	39	.424	.256
20	.625	.260	40	.647	.295

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 5, FORM C

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.189	.142	21	.600	.188
2	.135	.079	22	.530	.304
3	.108	.110	23	.636	.270
4	.227	.277	24	.524	.288
5	.308	.323	25	.686	.418
6	.135	.240	26	.665	.075
7	.135	.278	27	.708	.315
8	.243	.194	28	.557	.326
9	.297	.079	29	.763	.139
10	.308	.283	30	.638	.223
11	.346	.291	31	.730	.317
12	.324	.283	32	.730	.190
13	.422	.257	33	.789	.297
14	.411	.223	34	.757	.308
15	.481	.423	35	.746	.198
16	.378	.218	36	.800	.179
17	.524	.399	37	.897	.165
18	.486	.123	38	.886	.062
19	.395	.358	39	.892	.254
20	.573	.360	40	.886	.240

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 6, FORM A

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.887	.066	21	.760	.210
2	.949	.170	22	.618	.434
3	.887	.152	23	.407	.327
4	.927	.229	24	.178	.214
5	.956	.092	25	.633	.361
6	.905	.330	26	.593	.468
7	.920	.249	27	.727	.241
8	.960	.130	28	.371	.220
9	.949	.106	29	.749	.407
10	.869	.263	30	.775	.389
11	.524	.284	31	.796	.368
12	.331	.306	32	.196	.232
13	.218	.276	33	.764	.335
14	.575	.363	34	.430	.396
15	.465	.464	35	.084	.110
16	.309	.174	36	.393	.513
17	.705	.215	37	.415	.275
18	.302	.484	38	.618	.424
19	.527	.177	39	.607	.217
20	.302	.363	40	.360	.289

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 6, FORM B

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.258	.261	21	.640	.261
2	.210	.339	22	.918	.250
3	.592	-.072	23	.884	.136
4	.296	.126	24	.914	.056
5	.787	.262	25	.936	.146
6	.397	.389	26	.371	.205
7	.243	.178	27	.888	.268
8	.283	.352	28	.592	.316
9	.127	.278	29	.637	.389
10	.217	.199	30	.363	.328
11	.581	.285	31	.644	.304
12	.487	.102	32	.390	.443
13	.588	.185	33	.633	.358
14	.846	.245	34	.412	.187
15	.360	.177	35	.820	.231
16	.712	.406	36	.903	.148
17	.835	.272	37	.936	.164
18	.506	.334	38	.929	.275
19	.745	.467	39	.745	.384
20	.831	.225	40	.959	.147

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 7, FORM A

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.944	.147	21	.659	.392
2	.880	.181	22	.644	.347
3	.809	.188	23	.678	.454
4	.910	.327	24	.603	.401
5	.816	.253	25	.618	.365
6	.846	.316	26	.483	.445
7	.839	.183	27	.629	.306
8	.873	.176	28	.596	.249
9	.861	.270	29	.581	.345
10	.757	.337	30	.543	.271
11	.682	.428	31	.581	.411
12	.846	.369	32	.569	.282
13	.869	.336	33	.487	.229
14	.779	.377	34	.524	.303
15	.745	.283	35	.382	.349
16	.719	.324	36	.457	.305
17	.745	.270	37	.363	.253
18	.663	.304	38	.427	.316
19	.577	.415	39	.311	.257
20	.670	.283	40	.172	.315

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 7, FORM B

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.120	.183	21	.682	.355
2	.356	.286	22	.674	.397
3	.311	.173	23	.670	.499
4	.333	.197	24	.757	.448
5	.487	.360	25	.685	.364
6	.423	.406	26	.783	.360
7	.498	.297	27	.783	.412
8	.401	.175	28	.768	.291
9	.573	.367	29	.809	.381
10	.566	.288	30	.738	.390
11	.532	.280	31	.858	.211
12	.607	.366	32	.835	.356
13	.693	.261	33	.861	.338
14	.543	.273	34	.854	.086
15	.637	.395	35	.861	.313
16	.577	.296	36	.843	.311
17	.663	.502	37	.910	.373
18	.596	.377	38	.835	.357
19	.655	.506	39	.869	.269
20	.734	.501	40	.948	.238

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 8, FORM A

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.922	.103	21	.531	.191
2	.964	.237	22	.620	.359
3	.870	.440	23	.573	.227
4	.859	.295	24	.708	.375
5	.896	.347	25	.578	.495
6	.849	.146	26	.583	.464
7	.828	.330	27	.510	.335
8	.885	.205	28	.474	.205
9	.865	.257	29	.469	.409
10	.771	.494	30	.271	.326
11	.792	.251	31	.292	.236
12	.802	.119	32	.411	.257
13	.849	.273	33	.297	.195
14	.818	.465	34	.339	.194
15	.589	.207	35	.229	.269
16	.807	.291	36	.229	.274
17	.646	.275	37	.245	.278
18	.672	.467	38	.083	.039
19	.630	.439	39	.083	.074
20	.708	.439	40	.135	.068

DIFFICULTY AND DISCRIMINATION VALUES
OBTAINED FROM QUIZ 8, FORM B

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.126	.244	21	.618	.485
2	.173	.269	22	.733	.304
3	.141	-.002	23	.702	.234
4	.204	.271	24	.597	.269
5	.251	.478	25	.832	.275
6	.325	.329	26	.670	.325
7	.372	.245	27	.806	.405
8	.346	.322	28	.838	.321
9	.403	.471	29	.785	.288
10	.298	.245	30	.791	.282
11	.356	.409	31	.885	.299
12	.429	.293	32	.812	.234
13	.524	.359	33	.906	.226
14	.545	.294	34	.890	.281
15	.576	.355	35	.869	.089
16	.471	.523	36	.890	.339
17	.607	.372	37	.906	.222
18	.571	.217	38	.895	.196
19	.618	.235	39	.958	.251
20	.524	.233	40	.927	.075

DIFFICULTY AND DISCRIMINATION VALUES OBTAINED FOR QUIZ 8,
FORM A FOR STUDENTS USING KNOWLEDGE OF RESULTS
ANSWER BOARDS

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.915	.378	21	.644	.425
2	.932	.195	22	.712	.443
3	.831	.477	23	.559	.419
4	.763	.069	24	.610	.347
5	.898	.153	25	.492	.347
6	.932	.218	26	.644	.281
7	.881	.460	27	.407	.392
8	.881	.291	28	.559	.536
9	.932	.333	29	.475	.281
10	.797	.402	30	.339	.249
11	.780	.137	31	.305	.331
12	.831	.178	32	.492	.359
13	.763	.509	33	.441	.707
14	.814	.242	34	.356	.453
15	.627	.359	35	.220	.497
16	.847	.181	36	.203	.429
17	.576	.333	37	.322	.310
18	.610	.205	38	.288	.365
19	.610	.371	39	.220	.566
20	.661	.165	40	.271	.304

DIFFICULTY AND DISCRIMINATION VALUES OBTAINED FOR QUIZ 8,
FORM B FOR STUDENTS USING KNOWLEDGE OF RESULTS
ANSWER BOARDS

Item Number	Diffi- culty	Discrimi- nation	Item Number	Diffi- culty	Discrimi- nation
1	.150	.243	21	.633	.330
2	.167	.478	22	.750	.418
3	.167	.300	23	.733	.302
4	.283	.248	24	.700	.334
5	.217	.562	25	.800	.320
6	.267	.312	26	.683	.425
7	.467	.330	27	.867	.305
8	.433	.230	28	.833	.103
9	.433	.460	29	.817	.070
10	.500	.127	30	.783	.454
11	.367	.257	31	.933	.200
12	.533	.222	32	.900	.280
13	.583	.244	33	.933	.333
14	.433	.460	34	.833	.232
15	.617	.434	35	.967	.108
16	.650	.509	36	.933	.035
17	.767	.366	37	.917	.342
18	.733	.178	38	.883	.229
19	.833	.034	39	1.000	.000
20	.717	.045	40	.933	.097

APPENDIX B

TEST STATISTICS FOR QUIZZES 1, 2, 3, AND 4

THE MEANS, STANDARD DEVIATIONS, AND
POPULATION SIZE FOR QUIZZES
1, 2, 3, AND 4

Form	Mean	Standard Deviation	Number
1A	23.486	3.950	319
1B	20.403	4.068	325
2A	23.716	4.219	320
2B	22.542	4.279	319
3A	26.253	4.288	312
3B	26.137	5.036	313
4A	23.950	5.379	299
4B	25.141	4.820	305

APPENDIX C

MEANS OF STUDENTS' ESTIMATES OF DIFFICULTY
FOR THE ITEMS IN QUIZZES 7

MEANS OF STUDENTS' ESTIMATES OF DIFFICULTY
FOR THE ITEMS IN QUIZ 7 FORM A

Total Score Division			Total Score Division		
Item	Upper Third	Lower Third	Item	Upper Third	Lower Third
1	1.99	2.43	21	2.63	3.18
2	2.65	2.98	22	3.13	3.23
3	3.05	3.20	23	3.12	3.24
4	2.73	3.16	24	2.76	3.15
5	3.26	3.27	25	3.79	3.55
6	3.18	3.12	26	3.22	3.66
7	3.35	3.16	27	3.56	3.60
8	3.06	3.09	28	3.58	3.51
9	2.56	2.60	29	3.33	3.50
10	3.30	3.34	30	3.76	3.71
11	2.76	3.09	31	3.03	3.54
12	2.99	3.09	32	3.64	3.30
13	3.20	3.52	33	3.64	3.46
14	3.01	3.27	34	3.08	3.09
15	3.15	3.18	35	2.84	2.99
16	3.03	3.20	36	4.13	4.04
17	3.28	3.09	37	3.50	3.62
18	2.76	3.10	38	3.85	3.86
19	3.42	3.59	39	3.97	3.73
20	2.56	3.44	40	4.15	4.03

MEANS OF THE STUDENTS' ESTIMATES OF DIFFICULTY
FOR THE ITEMS IN QUIZ 7, FORM B

Total Score Division			Total Score Division		
Item	Upper Third	Lower Third	Item	Upper Third	Lower Third
1	3.81	3.51	21	2.58	3.05
2	3.67	3.38	22	3.31	3.44
3	3.70	3.51	23	2.57	2.78
4	3.22	3.19	24	2.94	2.96
5	3.38	3.45	25	2.94	2.87
6	2.73	2.59	26	3.07	3.00
7	2.48	2.79	27	3.18	3.30
8	3.25	3.05	28	3.17	3.28
9	3.09	2.98	29	3.03	3.04
10	2.72	3.16	30	2.50	2.75
11	3.52	3.38	31	3.00	2.84
12	2.91	3.22	32	2.42	2.43
13	3.07	3.00	33	2.83	2.91
14	3.07	3.31	34	3.20	2.98
15	2.86	3.22	35	2.89	2.83
16	3.34	3.40	36	2.82	3.00
17	2.57	2.65	37	2.46	3.06
18	3.20	3.23	38	3.03	3.15
19	2.58	2.93	39	2.61	3.25
20	2.50	2.89	40	2.26	2.65

BIBLIOGRAPHY

1. Anastasi, Anne. Psychological testing. New York: Macmillan, 1954.
2. Bruner, J. S., & Postman, L. On the perception of incongruity: a paradigm. J. Personal., 1949, 18, 206-223.
3. Bruner, J. S., Postman, L., & Rodrigues, J. Expectation and the perception of color. Amer. J. Psychol., 1951, 64, 216-227.
4. Capron, Virginia L. The relative effect of three orders of arrangement of items upon pupils' scores in certain arithmetic and spelling tests. J. Ed. Psychol., 1933, 24, 687-695.
5. Carter, H. How reliable are the common measures of difficulty and validity of objective test items? J. Psychol., 1942, 13, 31-39.
6. Cronbach, L. J. Essentials of psychological testing. (2nd ed.) New York: Harper, 1960.
7. Davis, F. B. Item selection techniques. In E. F. Lindquist (ed.) Educational measurement. Washington: American Council on Education, 1951, pp. 266-328.
8. Davis, F. B. Notes on test construction: the reliability of item-analysis data. J. Ed. Psychol., 1946, 37, 335-390.
9. Edwards, A. L. Statistical methods for the behavioral sciences. New York: Rinehart, 1954.
10. Gibbons, C. C. The predictive value of the most valid items of an examination. J. Ed. Psychol., 31, 1940, 615-621.
11. Greene, H. A., Jorgensen, A. N., & Gerberich, J. R. Measurement and evaluation in the secondary school. New York: Longmans, Green, 1943.
12. Hilgard, E. R., & Jandron, E. Stability of item statistics in two institutions. J. Ed. Psychol., 1960, 51, 195-198.
13. Jahoda, Marie, Deutsch, M., & Cook, S. W. Research methods in social relations, part two. New York: Dryden Press, 1951.
14. Kroll, A. Item validity as a factor in test validity. J. Ed. Psychol., 1940, 31, 425-436.

15. Kuder, G. F., & Richardson, M. W. The theory of estimation of test reliability. Psychometrika, 1937, 2, 151-160.
16. Longstaff, H. P. Fakability of the Strong Interest Blank and the Kuder Preference Record. J. appl. Psychol., 1948, 32, 360-369.
17. Longstaff, H. P. & Jurgensen, C. E. Fakability of the Jurgensen Classification Inventory. J. appl. Psychol., 1953, 37, 86-99.
18. Lorge, I., & Kruglov, Lorraine. The improvement of estimates of test difficulty. Educ. Psychol. Measmt., 1953, 13, 34-46.
19. Lund, K. Test performance as related to order of item difficulty, anxiety and intelligence. Unpublished doctoral dissertation, Northwestern Univer., 1953.
20. Mensh, I. N. The effects of test abbreviation upon responses to the individual items. J. Clin. Psychol., 1949, 5, 22-36.
21. Murphy, G. Personality. New York: Harper, 1947.
22. Pressey, S. L., Robinson, F. P., & Horrocks, J. E. Psychology in education. New York: Harper, 1959.
23. Remmers, H. H. & Gage, N. L. Educational measurement and evaluation. New York: Harper, 1943.
24. Ross, C. C. Measurement in today's schools. (2nd ed.) New York: Prentice-Hall, 1947.
25. Ruch, G. M. The objective or new-type examination. New York: Scott, Foresman, 1929.
26. Siipola, E. M. A group study of some effects of preparatory set. Psychol. Monogr., 1935, 46, No. 210.
27. Thorndike, R. L., & Hagen, Elizabeth. Measurement and evaluation in psychology and education. (2nd ed.) New York: Wiley, 1961.
28. Tinkelman, S. Difficulty prediction of test items. Teachers College Contributions to Education, No. 941. New York: Bureau of Publications, Teachers College, Columbia University, 1947.

AUTOBIOGRAPHY

I, Marshall Hallock Brenner, was born in Kalamazoo, Michigan, May 27, 1933. I received my secondary-school education at University High in Kalamazoo and then attended Kalamazoo College, where I received the Bachelor of Arts degree in 1955. I received the Master of Arts degree from Bowling Green State University, Bowling Green, Ohio, in January, 1959. In the fall of 1959 I entered the Graduate School of The Ohio State University for study toward the Ph.D. degree in the Department of Psychology.