

**RESEARCH
BULLETIN**

**A THEORY OF TEST SCORES AND
THEIR RELATION TO THE TRAIT MEASURED**

Frederic M. Lord

(Ph.D. Thesis, Princeton University, 1951)

Educational Testing Service

April 27, 1951

TABLE OF CONTENTS

	<u>Page</u>
I. INTRODUCTION	1
II. THEORETICAL TREATMENT	5
A. <u>Restrictions, Assumptions, and Definitions</u>	5
1. Restrictions on the Scoring Method	5
2. Definition of "Ability"	5
3. Restriction on the Population of Examinees	8
4. Restriction on the Nature of the Test	10
5. Alternative Assumptions and Restrictions	11
B. <u>The Bivariate Distribution of Test Score and Ability</u>	12
1. The Frequency Distribution of Scores for Examinees at a Given Level of Ability	12
2. The Bivariate Frequency Distribution of Test Score and Ability	13
3. The Regression of Test Score on Ability	14
4. The Regression of Ability on Test Score	16
5. The Standard Error of Measurement at a Specified Ability Level	16
6. The Product Moment Correlation between Test Score and Ability	20
7. The Curvilinear Correlation of Test Score on Criterion Score; The Test Reliability	21
8. The Magnitude of the Practical Effect of the Curvilinear Regression	25
C. <u>The Discriminating Power of the Test at a Given Level of Ability</u>	29
1. Derivation of an Index of Discriminating Power	29
2. The Conditions for Maximum Discrimination	33

	<u>Page</u>
C. 3. Numerical Illustrations of the Discrimination Index	35
4. Relation of the Discrimination Index to the Test Reliability	38
D. <u>The Frequency Distribution of Test Scores</u>	39
E. <u>The Limiting Frequency Distribution of Test Scores for Large n -- The Frequency Distribution of True Scores</u>	45
1. Derivation of the Distribution	45
2. Illustrative Examples of the Distribution of Relative True Scores	48
F. <u>The Bivariate Distribution of Scores on Two Tests Measuring the Same Ability</u>	49
G. <u>Extension of Results to Multiple Choice Items</u>	52
III. <u>EMPIRICAL VERIFICATION</u>	53
H. <u>The Plan</u>	53
I. <u>The Data</u>	54
J. <u>Procedure</u>	56
1. Calculating the Item Statistics	56
2. Selection of Tests	59
3. Obtaining the Theoretical Bivariate Frequency Distribution of Test Scores and Ability	60
4. Obtaining the Theoretical Univariate Distributions of Test Scores	61
5. Obtaining the Theoretical Bivariate Frequency Distribution for Two Subtest Scores	62
K. <u>Comparison of Theoretical and Actual Results</u>	63
1. Comparison of Univariate Frequency Distributions of Test Score	63
2. Comparison of Bivariate Frequency Distributions of Test Score	69

	<u>Page</u>
IV. SUMMARY AND CONCLUSIONS	71
REFERENCES	76
APPENDICES	80
Figures 4-18	81
Tables 4-37	93
Summary of Notation (Appendix A)	1A

LIST OF FIGURES

<u>Figure</u>		<u>Following Page</u>
1	Item Characteristic Curve When $h_1 = -.562$ and $R_1 = .531$	10
2	Discrimination Index and Standard Error of Measurement at Various Levels of Ability, and Also Regression of Test Score on Ability, for Four Hypothetical 100-Item Tests.	36
2A	The Discrimination Index as a Function of Ability, for Each of Five Tests with Specified Values of σ_h^2	36
3	Frequency Distributions of Relative Scores on Four Infinitely Long Tests Composed of Equivalent Items of 50 Per Cent Difficulty Whose Correlations with Ability Are As Indicated.	48
4-18	(Appendix).	81 ff.

LIST OF TABLES

<u>Table</u>		<u>Page</u>
1	Standard Error of the Tetrachoric Correlation between Two Items, for Specified Combinations of Item Difficulties, When the True Correlation Is .30 and the Number of Cases Is 136.	58
2	Probability That Certain Values of Chi-Square Having the Specified Degrees of Freedom, Will Be Exceeded in Random Sampling.	68
3	Probability That Certain Values of Chi-Square Having the Specified Degrees of Freedom, Will Be Exceeded in Random Sampling.	70
4-37	(Appendix).	93 ff.

**A THEORY OF TEST SCORES AND
THEIR RELATION TO THE TRAIT MEASURED**

I. INTRODUCTION *

A mental trait of an examinee is commonly measured in terms of a test score that is a function of the examinee's responses to a group of test items. For convenience we will speak here of the "ability" measured by the test, although our conclusions will apply to many tests that measure mental traits other than those properly spoken of as "abilities."

The ability itself is not a directly observable variable, and hence its magnitude, in terms of whatever metric we may choose to adopt, can only be inferred from the examinee's responses to the test items. The test score most commonly used as a measure of ability is the sum of the item scores when each response is scored 0 or 1. The metric provided by such a score has the serious disadvantage that it is largely a function of the particular characteristics of the items that happen to compose the particular test administered. An illustration of this statement is seen in the fact that two tests of the same ability administered to the same group of examinees may yield two score distributions of entirely different shapes. If one of these distributions were positively skewed and the other were negatively skewed, for example, one might be tempted to conclude that the group was composed predominantly of incompetent individuals in the former case, or of highly competent individuals in the latter case. As shown by this illustration, it is not possible to consider the test score as simply "ability" plus or minus an independent, normally distributed error of measurement.

- - - - -
*The author is greatly indebted to Dr. Harold Gulliksen and Dr. Ledyard R Tucker for their advice and criticism. A number of the developments presented have arisen from the suggestions of Dr. Tucker.

It would be desirable to define as a measure of ability some function of the item scores that will remain invariant for any examinee, even though the items composing the test are changed. Because of the inevitable presence of errors of measurement, it is of course impossible to determine with complete precision such an invariant measure of ability for a given examinee from data on a finite number of test items. In spite of this indeterminacy, it is nevertheless possible under certain conditions to define a metric for ability such that the frequency distribution of ability in the group tested will remain the same even though the composition of the test is changed. Furthermore, the bivariate frequency distribution of test score and ability will be completely determined as a function of the usual item statistics. If we are willing to accept this invariant metric as a useful metric for describing the underlying ability measured, and if in any given case the actual data are found upon examination to meet the necessary theoretical conditions, we have the result that all the properties of the test score in relation to the underlying ability (and in addition the univariate frequency distribution of test score and the bivariate distribution of scores on two tests measuring the same ability) will thus have been expressed as functions of the usual item statistics.

In this way we can attempt to answer such questions as the following: Under what circumstances, if any, can we expect platykurtic, rectangular, or U-shaped distributions of test scores? Do equal units of test score correspond to equal units of ability, as here defined? How does the standard error of measurement vary with the ability level of the examinees tested? How much do examinees having a given test score vary in ability?

What is the discriminating power of the test for examinees at various specified ability levels? How does the possibility of guessing the correct answers to the test items affect the properties of the test score? In general, how do changes in item difficulty or item intercorrelation affect the answers to these questions, and how can we select the items that compose the test in such a way as to achieve any desired result?

The present article attempts to develop a theory of test scores that will throw light on these and other questions. The correspondence of certain of the theoretical results with observed results on actual test data is checked in an empirical study reported in Part III.

To the writer's best knowledge, the only published work comparable to the present development is that of Lawley (19, 20), who, however treats only the special case where 1) certain item difficulty indices have a normal frequency distribution for the items composing the test, 2) the item intercorrelations are all equal, and 3) the items cannot be answered correctly by guessing. Lawley does not deal with the frequency distribution of test scores for the total group of examinees, nor with certain other problems treated here. Although actually derived independently, the present development may be considered as a generalization and extension of Lawley's formulations, removing the three restrictions listed and developing certain new theoretical conclusions.

More recently, Carroll (2) has made a similar theoretical development in which he assumes that the probability that an examinee will answer an item correctly is a linear function of the examinee's ability (whereas Lawley and the present writer made the more usual assumption that this

probability is a normal ogive function of ability). Carroll requires that all items have the same correlation with the trait measured, but he does not restrict the distribution of item difficulties, nor does he rule out the case where the items can be answered correctly by guessing. Any more detailed discussion of his work must await its completion and publication.

The assumptions that we will use here have also previously been used with considerable success by Tucker (36) to predict how different groups of examinees will perform on an item. Tucker has also (35) used similar but more restrictive assumptions in a theoretical study of the effect on test validity of changes in the intercorrelation of the items. Brogden (1) has used similar assumptions in a theoretical and numerical study of the relation of item intercorrelation and of item difficulty distributions to test validity and reliability.

II. THEORETICAL TREATMENT

A. Restrictions, Assumptions, and Definitions

1. Restrictions on the Scoring Method

Consideration will be restricted to the situation where the examinee attempts every item in the test, and the responses are all scored either 0 or 1. Let x_i ($i = 1, 2, \dots, n$) represent the score assigned to item i ; so x_i is a dichotomous variable that can assume only the values 1 or 0. It will be convenient to use the language of achievement testing and to speak of these alternatives as corresponding to "correct" and "incorrect" item responses, respectively.

We will further restrict consideration to the case where the test score (s) is the sum of the scores on the n items of which the test is composed:

$$s = \sum_{i=1}^n x_i . \quad (A1.1)$$

Here and elsewhere all summations are over the item subscripts i, j ($= 1, 2, \dots, n$) unless otherwise specified. (A summary of the notation is provided in the Appendix for easy reference.)

2. Definition of "Ability"

For present purposes, any useful definition of the underlying ability measured by the test must be in terms of observable variables. Since, in the final analysis, the only observable variables under consideration are the item responses, any operational definition of ability, for present purposes, must consist of a statement of a relationship between ability and item responses.

The relationship to be used here for this purpose may be stated as follows: the probability that an examinee will answer an item correctly is a normal ogive function of his ability. Denoting this probability for the i -th item by P_i , we may state this relationship more explicitly:

$$P_i = \int_{-\infty}^{(c-a_i)/b_i} N(y) dy, \quad (A2.1)$$

where c is the measure of ability, a_i and b_i are values characterizing the item, y is simply a variable of integration, and $N(y)$ is the normal frequency function

$$N(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}. \quad (A2.2)$$

We will use the notation $A(y_0)$ to indicate the area of the standardized normal curve lying above the given point, y_0 . Thus, because of the symmetry of the normal curve, we may write

$$P_i = A\left(\frac{a_i - c}{b_i}\right). \quad (A2.3)$$

When P_i is plotted as a function of c , we obtain the curve that Tucker has called the item characteristic curve, and that Lazarsfeld (21) and others have called the trace line of the item.

The relationship expressed by A2.1 has been widely used in psychophysics in other connections and has in recent years been used to represent the probability of answering a test item correctly by Guilford (10),

Richardson (31), Mosier (25, 26), Ferguson (6), Lawley (op. cit.), Lorr (23), Tucker (op. cit.), and others. This relationship, indeed, is implicitly assumed whenever a tetrachoric correlation is calculated between a test item and a dichotomized, normally distributed measure of ability.

In order to utilize Equation A2.1, we must first determine the values of a_i and b_i for each item. If we make no further assumptions, this can be done approximately, in actual practice, by the following approach, suggested by Tucker: 1) Start with a test consisting of a very large number of items all presumably measuring the same ability (c). 2) Obtain the test score of each individual in a very large group of examinees. 3) Assume that this score is highly related to c (not necessarily linearly) so that examinees having identical scores may be treated for practical purposes as having identical values of c . 4) Record for each item the percentage of examinees at each given score level who answer the item correctly; these percentages may conveniently be considered as plotted as a function of the score, thus constituting the item characteristic curve. 5) If possible, transform the score scale until all the item characteristic curves are normal ogives, to a close approximation. 6) Determine the values of a_i and b_i by standard methods of probit analysis (Guilford, op. cit., p. 173, and especially Finney (8) who presents recently developed maximum likelihood methods for dealing with a variety of problems in this area). If Step 5 is possible, then this procedure will define a measure of ability for our present purpose; if Step 5 is impossible, the theory outlined in the present paper will be inapplicable.

Further attention will be given to the item characteristic curve in the next section, after we have discussed a further restriction on the data to be considered.

3. Restriction on the Population of Examinees

First let us restrict consideration to a sample of examinees so large that sampling problems need not enter into the discussion. Such problems are of practical importance, but they would unnecessarily complicate the theoretical picture.

If ability has been empirically defined for any given set of data by the six steps outlined in the preceding section, the frequency distribution of ability in the group tested will have been determined by this procedure. If a mathematical frequency curve can be fitted to this distribution, it is possible to go ahead with the theoretical development along the lines to be presented. It will be more profitable, however, for us to restrict our attention to the case where ability is normally distributed in the group of examinees tested. We may, without loss of generality, choose the scale of measurement for ability so that the mean ($\underline{M}$) and standard deviation ($\underline{\sigma}$) of $\underline{c}$ are respectively 0 and 1:

$$M_c = 0, \quad \sigma_c = 1. \quad (A3.1)$$

In spite of the restriction made in the preceding paragraph, many of the formulas to be obtained will nevertheless be valid for the general case when no restriction is placed on the distribution of ability in the group tested. Attention will be called to the generality of these formulas when they are presented.

The restriction that we have placed on the character of the group tested will simplify the determination of the values of a_i and b_i for each item. Tucker (35, Equations 10 and 14) has given the formulas necessary for expressing a_i and b_i in terms of the item difficulty, p_i (percentage of correct answers in the total group of examinees), and of the biserial correlation, R_i , between item response and ability. In our notation, his results are:

$$R_i = \frac{1}{\sqrt{1 + b_i^2}} ; \quad (A3.2)$$

$$p_i = A(h_i) , \quad (A3.3)$$

where $h_i = a_i R_i$.

We will assume henceforth, unless otherwise specified, that $R_i^2 < 1$.

We may then write

$$K_i = \sqrt{1 - R_i^2} , \quad (A3.4)$$

$$g_i = \frac{h_i - R_i c}{K_i} . \quad (A3.5)$$

Let us use A3.2 and A3.3 to eliminate a_i and b_i from the equation for the item characteristic curve (A2.3) so as to obtain the following formula, which will be most convenient for our further development:

$$P_i = A(g_i) . \quad (A3.6)$$

It should be borne in mind that both g_i and P_i are functions of c .

Figure 1 is presented to provide an illustration of an item characteristic curve. The curve, calculated from A3.6, represents an item for which $h_1 = -.562$ and $R_1 = .531$ (Item 44 of Part III).

The actual numerical value of h_i can be readily obtained for any given set of data from the item difficulty, by means of Equation A3.3. An approximation to the value of R_i may usually be obtained from the biserial correlation of the item with score on a very long test of the ability under consideration, after correcting this biserial correlation for attenuation due to the unreliability of the test. Thus the numerical values of the item characteristic curves may be determined in practice from appropriate empirical data without the use of probit analysis methods.

A more accurate method of determining R_i will be discussed in the following section.

4. Restriction on the Nature of the Test

In order to utilize the item characteristic curve for much further development, it is necessary to make some assumption concerning the probability that an examinee at a given level of ability will answer both of two items correctly. Here we will restrict consideration to tests composed of items such that the ability underlying the test is the only common factor in the matrix of tetrachoric item intercorrelations (see Wherry and Gaylord (37) for a discussion of the reasonableness of such a restriction). From this restriction it follows that, for fixed $\underline{c}$, P_i is independent of P_j , where i and j represent any two different test items.

Since $\underline{c}$ is the common factor in the matrix of tetrachoric item inter-

correlations, we see that R_i^2 is the communality of the i -th item in this matrix. The numerical values of R_i may therefore be determined directly from this matrix by factor analysis methods, and it is not necessary to have available a very long test whose scores provide an approximate measure of ability.

5. Alternative Assumptions and Restrictions

Before proceeding with our development, it will be of interest to point out, as Brogden (1, p. 202, footnote) has done, that there is another, mathematically equivalent set of assumptions that might well be used instead of those outlined in the three preceding sections.

Just as the response data on two items may be represented by a 2×2 table, such as is used in calculating tetrachoric correlations, so may the response data on n items be represented by a $2 \times 2 \times \dots \times 2$ table in n dimensions containing 2^n cells. The equivalent set of assumptions are as follows:

- a) The 2^n frequencies in the multivariate distribution of the item responses are such as could have arisen by the dichotomization of each of the variables in a normal multivariate population.
- b) The correlations between the variables in the normal multivariate population have only one common factor.

It is worth pointing out that it is possible (although excessively difficult in practice) to test whether or not any given set of data conform to these two requirements. The first of these two requirements cannot be contradicted by the data unless $n \geq 3$, the second unless $n \geq 4$.

B. The Bivariate Distribution of Test Score and Ability

1. The Frequency Distribution of Scores for Examinees at a Given Level of Ability

Since $\underline{c}$ is the only common factor of the tetrachoric item intercorrelations, it follows that P_i and P_j ($i \neq j$) are independent when $\underline{c}$ is fixed. We thus see, for example, that the probability of success on both items $\underline{i}$ and $\underline{j}$ is $P_i P_j$; that the probability of success on one and failure on the other is $P_i Q_j + Q_i P_j$, where $Q_i = 1 - P_i$; and finally that the probability of failure on both items is $Q_i Q_j$. Generalizing this result to $\underline{n}$ items, we find that the probability of success on $\underline{s}$ items out of $\underline{n}$, for examinees at a given level of ability, is given by the sum of the appropriate terms in the expansion of the product

$$\prod_{i=1}^n (Q_i + P_i) . \quad (\text{B1.1})$$

(It must be borne in mind here that P_i and Q_i are functions of $\underline{c}$, as shown by Equations A3.6 and A3.5.) Since the score is the number of items answered correctly, the terms of the expansion of B1.1 represent the distribution of test scores ($\underline{s}$) at a given level of ability. If we denote this conditional distribution of $\underline{s}$ by $f_{s.c}$, we obtain by expanding B1.1 the result that

$$f_{s.c} = \Sigma^* \prod_{s \text{ } i} P_i \prod_{n-s \text{ } i} Q_i , \quad (\text{B1.2})$$

where $\prod_{s \text{ } i} P_i$ is the product of the values of P_i for any $\underline{s}$ values of $\underline{i}$, $\prod_{n-s \text{ } i} Q_i$ is the product of the values of Q_i for the remaining $n - s$ values of $\underline{i}$, and Σ^* is the sum of the $\binom{n}{s} = \frac{n!}{s!(n-s)!}$ such possible products.

If all P_i are equal, B1.1 reduces to the familiar binomial expression, and B1.2 becomes

$$f_{s.c} = \binom{n}{s} P^s Q^{n-s} . \quad (B1.3)$$

The conditional distribution of s for fixed c is not dependent in any way on the distribution of ability in the particular group of examinees that we happen to have under consideration. Equations B1.1, B1.2, and B1.3 will remain valid whether or not the distribution of c is normal in the group tested. If c is not normally distributed, however, the values of R_i necessary for computing P_i and Q_i must be obtained by the method of Section A2, not by the method of Section A3 or of Section A4.

2. The Bivariate Frequency Distribution of Test Score and Ability

We are now in a position to achieve a fundamental objective -- to find the bivariate frequency distribution of test score and ability. The desired distribution will of course be (38, p. 17) the product of the conditional distribution of B1.2 and the marginal distribution of c :

$$f_{cs} = N(c) f_{s.c} , \quad (B2.1)$$

where $f_{s.c}$ is the expression given in Equation B1.2. It should be noted that, for given values of c and s , f_{cs} is a function only of the item difficulties and the tetrachoric item intercorrelations, and hence the values of f_{cs} may be calculated from actual item analysis data.

If we drop for a moment the requirement that the distribution of c shall be normal in the group of examinees under consideration and allow

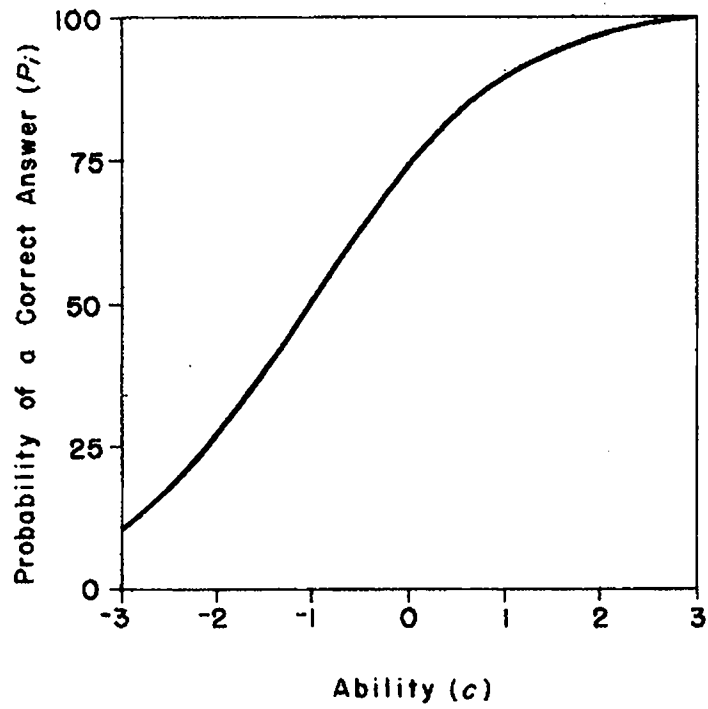


Figure 1. Item Characteristic Curve
When $h_1 = -.562$ and $R_1 = .531$.

the distribution to take any arbitrary form, f_c , the same line of argument that we have already used leads to a general formula for the bivariate distribution of ability and test score, as follows

$$f_{cs} = f_c \sum^* \prod_i P_i \prod_{n-s} Q_i . \quad (B2.2)$$

The values of R_i necessary for computing the expression under the summation sign in B2.2 must be obtained by the method of Section A2 when f_c is not normal.

The reader may wish to refer at this point to Tables 8 through 15, which provide numerical examples of the bivariate distribution of test score and ability.

3. The Regression of Test Score on Ability

Since the probability that an examinee at a given level of ability will answer a certain item correctly is denoted by P_i , his expected number of correct answers on an n -item test is $\sum P_i$. Rephrasing this statement, we may say that the average score of examinees at a given level of ability is

$$M_{s.c} = \sum P_i . \quad (B3.1)$$

Equation B3.1 is the usual expression for the mean of the distribution given by B1.2 (16, Vol. I, p. 122). Remembering that P_i is a function of c , we see that Equation B3.1 is the equation for the regression of test score on ability. Like the conditional distribution of s in B1.2, the equation for the regression curve is valid even if c is not normally

distributed in the group of examinees tested.

Figure 2 may be referred to at this point as providing illustrative examples of the regression of test score on ability for four hypothetical tests. The curvilinearity of this regression has been previously deduced by Brogden (1, p. 207).

Obviously the regression curve must be the sum of the n item characteristic curves -- in this case, of n normal ogives. If all the items have equal difficulties and equal intercorrelations, we have, dropping the subscript i , $M_{g.c} = nP$, and the regression curve is directly proportional to the item characteristic curve, as represented by P .

The slope of the item characteristic curve, and hence of the regression curve, is always positive, but the curves are practically horizontal for extreme values of c . If an item is sufficiently highly correlated with ability, however, the item characteristic curve will be practically vertical in some part of the range; and this statement remains true no matter how ability is defined or what metric is used to measure it. Since the regression curve must be the sum of the item characteristic curves, we see that the regression is inevitably curvilinear; and in particular, that it will be very strongly curved if the items are all of equal difficulty and have sufficiently high intercorrelations.

To be sure, any given group of examinees may have a limited range of ability within which the regression might appear to be approximately rectilinear. A derivation indicating the extent to which the regression curve will appear to be rectilinear in any given group of examinees will be presented in Section B8.

4. The Regression of Ability on Test Score

The other regression curve -- the regression of ability on test score -- can be written down in the form of a definite integral if desired. The integration cannot be carried out, however, and the integral itself is not readily manipulable.

It will be seen in the following section that as the number of items in the test is increased without limit, the scatter of the cases about the regression of test score on ability will vanish. As the number of items is increased, therefore, the regression of ability on test score will approach the regression of test score on ability as a limit. Since the shape of the latter regression does not change if the number of items is increased without changing their average characteristics, it follows that the regression of ability on test score has an ogive shape, at least whenever the test contains a large number of items.

5. The Standard Error of Measurement at a Specified Ability Level

The standard error of measurement for examinees at a specified level of ability is the standard deviation of $f_{s.c}$, which can be shown to be (16, loc. cit.)

$$\sigma_{s.c} = \sqrt{\Sigma P_i Q_i} = \sqrt{n(M_P M_Q - \sigma_P^2)}, \quad (B5.1)$$

where

$$M_P = \frac{1}{n} \Sigma P_i, \quad M_Q = \frac{1}{n} \Sigma Q_i, \quad (B5.2)$$

and

$$\sigma_P^2 = \frac{1}{n} \Sigma P_i^2 - M_P^2. \quad (B5.3)$$

As in the case of the regression equation, the formulas for the standard error of measurement at any given ability level remain valid irrespective of the frequency distribution of ability in the group tested.

It may be seen from B5.1 that the standard error of measurement will be practically zero for extreme positive or negative values of c . This conclusion fits in with the fact that, for any given test, at least in theory, and usually in practice, there are always examinees whose ability is so low that the test is not discriminating for them, and other examinees whose ability is so high that the test is likewise not discriminating for them. These are the examinees who are practically sure to get zero scores on the one hand, or perfect scores on the other. Obviously, the standard error of the test scores is practically zero for such examinees. Furthermore, this conclusion would necessarily be reached irrespective of any assumptions that may have been made in the present paper.

It is thus seen that the standard error of measurement of the test scores is necessarily smallest for those examinees for whom the test is least discriminating. Although the average of the standard errors of measurement, averaged over all ability levels in the group of examinees tested, may be a useful measure of the discriminating power of the test for the group as a whole, the standard error of measurement corresponding to a given ability level is thus seen to be very far from being a suitable measure of the discriminating power of the test for examinees at or near this ability level. An appropriate index of discriminating power for this purpose will be developed in Section C.

Let us consider for a moment the "relative score" or proportion of correct answers, which we will denote by $\underline{z}$, where

$$z = \frac{s}{n} . \quad (B5.4)$$

Since the standard deviation of $\underline{z}$ will be $1/n$ times the standard deviation of $\underline{s}$, we see from B5.1 that the standard error of measurement for relative score at a given level of ability is

$$\sigma_{z.c} = \sqrt{\frac{1}{n} (M_P M_Q - \sigma_P^2)} . \quad (B5.5)$$

When n becomes infinite, the relative score ($\underline{z}$) becomes the "relative true score" which we will denote by $\underline{t}$, i.e.,

$$t = \lim_{n \rightarrow \infty} z = \lim_{n \rightarrow \infty} \frac{s}{n} . \quad (B5.6)$$

The true score for a given test is usually defined as the average score that would be obtained on an infinite number of "equivalent" forms of the test. The same definition (in terms of relative scores) must be used here if we wish to consider $\underline{t}$ not merely as the relative score on an infinitely long test, but more specifically as the relative true score corresponding to the actual relative score obtained on a specified test of finite length.

Letting $n = \infty$ in B5.5, we see that $\sigma_{t.c} = 0$. This result has been pointed out by Brogden (1) and others. We thus reach the important conclusion that true score and ability have a perfect curvilinear correlation. If the scale of measurement for the true scores is transformed so as to give the distribution of the transformed true scores the same shape as the

distribution of $\underline{c}$, then the transformed true scores will be identical with $\underline{c}$. In particular, if $\underline{c}$ is normally distributed in the group tested, the value of $\underline{c}$ for each examinee will be identically equal to his normalized true score. Although we do not have the examinee's true score in practice for this purpose, the actual test score may be substituted for the true score if the number of items is sufficiently large.

We are now in a position to see that $\underline{c}$ provides a scale of measurement for the underlying ability that may be considered to be invariant for any given group of examinees even though the difficulties and intercorrelations of the items in the tests administered to the group are changed, provided the common factor of the item intercorrelations remains the same. This invariance of $\underline{c}$ is seen in the fact that, except for errors of measurement, the same value of $\underline{c}$ will be obtained for each examinee in a given group of examinees, irrespective of the difficulties and intercorrelations of the test items, and irrespective of the corresponding differences in the shapes of the test score distributions, provided always that all tests are measures of the same underlying ability.

Since true score and ability have a perfect curvilinear correlation, it follows that the standard error of measurement at a given ability level is the same as the standard error of measurement at a given true score level:

$$\sigma_{s.c} = \sigma_{s.t} \quad (B5.7)$$

The standard deviation of $\underline{c}$ ($\sigma_{c.s}$) for individuals having a given test score can be expressed in terms of definite integrals derived from equation

$$\sigma_s = \sigma_{\Sigma x_i} = \sqrt{\sum_{i,j} \sigma_i \sigma_j r_{ij}} \quad , \quad (B6.4)$$

where r_{ij} is the product moment (fourfold point) correlation between x_i and x_j . If $i = j$, r_{ij} is taken to be 1; otherwise the value of r_{ij} may be calculated from r'_{ij} (or from the usual fourfold table) by means of the standard formula (5, p. 124), which in our notation is

$$r_{ij} = \frac{A_2(h_i, h_j; r'_{ij}) - p_i p_j}{\sqrt{p_i q_i p_j q_j}} \quad , \quad (B6.5)$$

where

$$A_2(h_i, h_j; r'_{ij}) = \int_{h_i}^{\infty} \int_{h_j}^{\infty} N_2(u, v; r'_{ij}) dv du \quad , \quad (B6.6)$$

where $N_2(u, v; r'_{ij})$ is the normal bivariate frequency function for standardized variables u and v whose intercorrelation is r'_{ij} . In the present case, $A_2(h_i, h_j; r'_{ij})$ is the proportion of examinees answering both items i and j correctly. Values of $A_2(u, v; r)$ are tabled by Pearson (28, Vol. II).

Using Equations B6.1 through B6.4, the correlation of test score with ability may be written

$$r_{cs} = \frac{\sum R_i N(h_i)}{\sqrt{\sum \sigma_i \sigma_j r_{ij}}} \quad . \quad (B6.7)$$

7. The Curvilinear Correlation of Test Score on Criterion Score; The Test Reliability

The correlation ratio of test score on ability is identical with the

curvilinear correlation calculated from the best fitting regression curve, since this curve passes through all the conditional means, $M_{s.c}$. By definition this correlation ratio is equal to (3, p. 280)

$$\eta_{sc} = \sqrt{1 - \frac{E(\sigma_{s.c}^2)}{\sigma_s^2}}, \quad (B7.1)$$

where $E(\sigma_{s.c}^2)$ denotes the expected or average value of $\sigma_{s.c}^2$ for the total group of examinees. Now, $\sigma_{s.c} = \sigma_{s.t}$ (B5.7); and the average value of $\sigma_{s.t}^2$ is the square of the statistic that in test theory is commonly called the "standard error of measurement" ($S.E._{meas.}$). It is well known that

$$S.E._{meas.} = \sigma_s \sqrt{1 - r_{ss}}, \quad (B7.2)$$

where r_{ss} is the test reliability. Substituting B7.2 in B7.1, we thus have the interesting result that the curvilinear correlation of test score on ability is equal to the square root of the test reliability:

$$\eta_{sc} = \sqrt{r_{ss}}. \quad (B7.3)$$

This result is analogous to the well known fact that $r_{st} = \sqrt{r_{ss}}$.

We may consider the test reliability as the correlation of the actual test with a second, hypothetical, "equivalent" form. If we denote the score on the j -th item of the hypothetical equivalent test by X_j so that $X_j = 0$ or 1 , and if we assume that the two test scores have equal standard deviations, we have, by the usual formula for the correlation of sums,

$$r_{ss} = r(\Sigma x_i)(\Sigma x_j)$$

$$= \frac{\sum_{ij} \sigma_{x_i} \sigma_{x_j} r_{x_i x_j}}{\sigma_s^2} . \quad (B7.4)$$

Equation B7.4 is the same as the basic equation from which Kuder and Richardson (18, Equation 2) derive their reliability formula.

Since the equivalent test has only a hypothetical existence, the values of σ_{x_j} and $r_{x_i x_j}$ are known only by virtue of the definition of equivalence. According to this definition

$$\sigma_{x_j} = \sigma_j , \quad (B7.5)$$

the observed variance of the j -th item (see Equation B6.2); and

$$r_{x_i x_j} = r_{ij} \quad (i \neq j), \quad (B7.6)$$

the observed fourfold point correlation (see Equation B6.5). The value of $r_{x_i x_i}$ for our purposes may be determined from the requirement that the corresponding tetrachoric correlation, $r'_{x_i x_i}$, shall be such that the correlation matrix whose elements are $r'_{ij} = r'_{x_i x_j}$ shall have unit rank. We see from the definition of R_i that

$$r'_{x_i x_i} = R_i^2 ; \quad (B7.7)$$

hence, by B6.2 and B6.5,

$$r_{x_i x_i} = \frac{A_2(h_i, h_i; R_i^2) - p_i^2}{\sigma_i^2} . \quad (B7.8)$$

From Equations B6.4 and B7.3 through B7.8, we find the result that

$$\begin{aligned}
 r_{ss} = \eta_{sc}^2 &= \frac{\sum_1 A_2(h_1, h_1; R_1^2) - \sum_1 p_1^2 + \sum_{i \neq j} \sum_1 \sigma_1 \sigma_j r_{1j}}{\sigma_s^2} \\
 &= 1 - \frac{\sum_1 p_1 - \sum_1 A_2(h_1, h_1; R_1^2)}{\sigma_s^2} . \quad (B7.9)
 \end{aligned}$$

The expression given for the squared curvilinear correlation coefficient in B7.9 may be verified, if desired, by evaluating by direct integration the expected value (38, p. 29) of $\sigma_{s.c}^2$. Using B5.1, it will be found that

$$E(\sigma_{s.c}^2) = \int_{-\infty}^{\infty} (\sum_1 p_1 Q_1) N(c) dc = \sum_1 p_1 - \sum_1 A_2(h_1, h_1; R_1^2) . \quad (B7.10)$$

Substitution of this result in Equation B7.1 yields the same expression for the curvilinear correlation coefficient as that given by B7.9.

An examination of Equation B7.9 shows that any increase in R_1^2 , the communality of the items, will be accompanied by an increase in the curvilinear correlation, η_{sc} . This result is of special interest in view of Tucker's demonstration (35) and Brogden's later results (1) showing that a progressive increase in the item intercorrelations will lead at a rather early stage to a progressive decrease in the product moment correlation between test score and ability. This unexpected behavior of the product moment correlation is the result of the fact that as the item intercorrelations increase, the regression becomes more and more curvi-

linear, so that a straight line provides a progressively poorer fit, even though the scatter of the cases about the curved regression line is being continually reduced. Seen in this light, the situation does not seem to justify Brogden's inference that total score item analysis procedures are usually not worth while.

(No useful results for the other curvilinear correlation -- that of ability on test score -- have been obtained to date in view of the rather intractable expression obtained for $\sigma_{c.s.}$.)

8. The Magnitude of the Practical Effect of the Curvilinear Regression

As noted in Section 3, if the examinees tested all fall about a section of the regression curve that is practically linear, the essential curvilinearity of the regression has little practical effect. A measure of the magnitude of this effect is provided by a comparison of the curvilinear correlation of test scores on ability with the corresponding product moment correlation coefficient. This comparison will be made only for the case where the test is composed of equivalent items, in order to avoid excessive mathematical complications.

First we will obtain a formula for the product moment correlation for the case of equivalent items. From B6.4, dropping the subscripts i and j and remembering that $r_{ii} = 1$, we obtain

$$\sigma_s = \sqrt{npq + (n^2 - n) pqr} = \sqrt{[1 + (n - 1)r] npq} \quad . \quad (B8.1)$$

From B6.1, B6.2, and B8.1 we find the product moment correlation to be

$$r_{cs} = \frac{\sqrt{n} r_c}{\sqrt{1 + (n-1)r}} . \quad (\text{B8.2})$$

We next wish to have a formula for the curvilinear correlation, η_{sc} , for the case of equivalent items. In this case all values of r'_{ij} ($i \neq j$) are equal and may be denoted by a single symbol, r' . If the matrix of r'_{ij} is to have unit rank, the diagonal entries must also be equal to r' . Since the diagonal entries are the squares of the factor loadings, we have for this special case

$$R_1^2 = r' . \quad (\text{B8.3})$$

For equivalent items, dropping the subscripts i and j , Equation B6.5 becomes

$$r = \frac{A_2(h,h;r') - p^2}{pq} . \quad (\text{B8.4})$$

Replacing R_1^2 in B7.10 by r' and eliminating $A_2(h,h;r')$ from B7.10 by using B8.4, we obtain

$$E(\sigma_{s.c}^2) = npq(1 - r) . \quad (\text{B8.5})$$

From B7.1, B8.1, and B8.5 we thus obtain the result that for equivalent items

$$\eta_{sc}^2 = 1 - \frac{1 - r}{1 + (n-1)r} = \frac{nr}{1 + (n-1)r} . \quad (\text{B8.6})$$

This result will immediately be recognized as the Spearman-Brown formula

for predicting the reliability of a test composed of n equivalent items from the item intercorrelations, $\underline{r}$.

Since we are dealing with population parameters and not with sample statistics, there is no test of significance involved in comparing η_{sc} and r_{cs} . The most convenient procedure will be to examine the ratio η_{sc}^2/r_{cs}^2 . From B8.2 and B8.6 we have the very simple result that

$$\frac{\eta_{sc}^2}{r_{cs}^2} = \frac{r}{r_c^2} . \quad (B8.7)$$

By B6.3 and B8.3,

$$\frac{\eta_{sc}^2}{r_{cs}^2} = \frac{pqr}{R^2 N^2(h)} = \frac{pqr}{r' N^2(h)} . \quad (B8.8)$$

In order to obtain a clearer idea of the magnitude of this ratio, it will be helpful to expand $\underline{r}$ as a power series in r' . The usual power series used to determine a tetrachoric correlation coefficient from the frequencies in a fourfold table will serve our purpose (29, p. 369). For equivalent items, this series is

$$\frac{A_2(h, h; r') - p^2}{N^2(h)} = r' + \frac{1}{2}h^2 r'^2 + \frac{1}{6}(h^2 - 1)^2 r'^3 + \frac{1}{24}h^2(h^2 - 3)^2 r'^4 + \frac{1}{120}(h^4 - 6h^2 + 3)^2 r'^5 + \dots \quad (B8.9)$$

By B8.4, the left side of B8.9 is

$$\frac{A_2(h, h; r') - p^2}{N^2(h)} = \frac{pqr}{N^2(h)} \quad (\text{B8.10})$$

From B8.8, B8.9, and B8.10 we obtain

$$\begin{aligned} \frac{\eta_{sc}^2}{r_{cs}^2} = & 1 + \frac{1}{2}h^2r' + \frac{1}{6}(h^2 - 1)^2r'^2 + \frac{1}{24}h^2(h^2 - 3)^2r'^3 \\ & + \frac{1}{120}(h^4 - 6h^2 + 3)^2r'^4 + \dots \end{aligned} \quad (\text{B8.11})$$

If all items are of fifty per cent difficulty, $h = 0$, and we have

$$\frac{\eta_{sc}^2}{r_{cs}^2} = 1 + \frac{1}{6}r'^2 + \frac{3}{40}r'^4 + \dots \quad (\text{B8.12})$$

This last equation shows that when all items are of fifty per cent difficulty, η_{sc} will be practically equal to r_{cs} unless the tetrachoric item intercorrelations are much higher than is usual. In this case, therefore, the effect of the curvilinearity of the regression will be wholly negligible in actual practice. Equation B8.11, on the other hand, shows that whenever the item difficulties differ by a considerable extent from fifty per cent, -- when $h \geq 1$, say, -- η_{sc} may become considerably larger than r_{cs} , and the effect of the curvilinear regression may be quite noticeable in practice.

The reader may wish at this point to refer to Figure 2 and the accompanying text for numerical illustrations relating to these conclusions.

C. The Discriminating Power of the Test at a Given Level of Ability

1. Derivation of an Index of Discriminating Power

If we are going to use the test for selecting those individuals having the greatest amount of the ability measured by the test, we want a test that will discriminate accurately among examinees who are near the cutting score. If all selected examinees are to be treated alike and all rejected examinees are to be treated alike, irrespective of their test scores, then some measure of the discrimination of the test among examinees near the cutting score is the only correct measure of the validity of the test.

Most of the familiar statistics that suggest themselves as measures of this type of discrimination turn out on investigation to be wholly unsatisfactory. The correlation r_{cs} is of course unsuitable because it is only a measure of the average degree of discrimination over the entire group of examinees. The standard error of measurement at a given level of ability is usually thought of as a direct measure of the discriminating power of the test at that level; but we have already shown that the standard error of measurement for a given test at various levels of ability tends to be inversely related to the test's discriminating power.

A numerical example will serve to clarify the problem. Suppose almost all individuals for whom $c = -5$ obtain scores of either 0 or 1, and that their average score is 0.3. The standard error of measurement for these examinees is small -- perhaps 0.2 -- but it may still be that examinees for whom $c = -3$ obtain an average test score of only 0.31. In such a case it is obvious that the test does not discriminate appreciably between exam-

inees at the $c = -5$ level and examinees at the $c = -3$ level. Clearly any measure of discriminating power must take into account both the standard error of measurement and the slope of the regression.

Many different measures of discriminating power could be defined.

Let us start by comparing the distributions of test scores that will be obtained by examinees at two different ability levels, c_0 and c_1 . At each ability level the distribution of test scores will be approximately normal for large n , and if c_1 is close to c_0 these distributions will have approximately the same standard deviation. Under these conditions the amount of overlap of the two score distributions -- by overlap here is meant the extent to which the areas of the two frequency distributions coincide -- will vary directly as a function of the difference between the means ($M_{s.c_0}$ and $M_{s.c_1}$) of the distributions when this difference is expressed in standard deviation units. In other words, we may use as a measure of overlap the function

$$D' = \frac{|M_{s.c_0} - M_{s.c_1}|}{\sigma^*}, \quad (C1.1)$$

where the numerator is the absolute value of the difference between the means, and the denominator is some appropriate average of the standard deviations of the two distributions.

The function D' measures the distance between the two means. When c_1 is close to c_0 , this distance may be considered to vary in proportion to the distance $c_1 - c_0$. We are not interested in the distance between the two means corresponding to any specified distance $c_1 - c_0$, but rather in

the rate at which D' changes as a function of $c_1 - c_0$. For any given value $c = c_0$, this rate is

$$\begin{aligned} \lim_{c_1=c} \frac{|M_{s.c_1} - M_{s.c}|}{\sigma^*(c_1 - c)} &= \lim \frac{1}{\sigma^*} \cdot \lim \frac{|M_{s.c_1} - M_{s.c}|}{c_1 - c} \\ &= \frac{1}{\sigma_{s.c}} \frac{\partial M_{s.c}}{\partial c} . \end{aligned} \quad (C1.2)$$

Obtaining the indicated derivative from B3.1, A3.6, and A3.5, using the expression for $\sigma_{s.c}$ given by B5.1, and denoting the rate by $\underline{D}$, we have

$$D = \frac{1}{\sqrt{\Sigma P_1 Q_1}} \sum_1 \frac{R_1}{K_1} N(g_1) . \quad (C1.3)$$

Both P_1 and g_1 are functions of $\underline{c}$; so $\underline{D}$ is likewise a function of $\underline{c}$. We will use $\underline{D}$ as our measure of the discriminating power of the test at a specified level of ability (at a specified value of $\underline{c}$, not of $\underline{s}$). It is seen that $\underline{D}$ is the ratio of the slope of the regression curve to the standard deviation of the test scores at a fixed level of ability. The standard deviation is always positive, and we will assume that the slope of the regression will always be nonnegative. It is seen that the discrimination index will be zero when there is no discrimination, and that the more the discrimination, the higher the index, there being no upper limit to its possible value. If the item difficulties and the values of R_1 are available for any group of items meeting the assumptions made, the value of $\underline{D}$ can be calculated from equation C1.3 without excessive difficulty.

We may note that $\underline{D}$, like $M_{s.c}$ and $\sigma_{s.c}$ from which it is derived, is completely independent of the distribution of ability in the group tested. This is an advantage when a general description of the test is desired without reference to any particular group of examinees; it is a disadvantage if the effective discrimination of the test for a specified group of examinees is desired. Lawley (19) has derived an index of the effective discrimination of a test at different levels of ability for a specified group of examinees. The difference between "discriminating power" and "effective discrimination," as the terms are used here, may be illustrated as follows: A test may have low discriminating power for examinees in a certain range of ability; however, if in any given group of examinees there are only a few individuals spread out thinly over this range of ability, the rank order of these individuals on ability may be more accurately determined by the test scores than is the rank order of examinees in some other range of ability where the discriminating power (Cl.3) of the test is greater, but where there are many examinees of almost identical ability. The effective discrimination is greatest where the rank order of the examinees is most accurately determined.

Both Lawley's index and the one presented here are measures of discrimination at a specified ability level. For most practical purposes, it would be better to have an index measuring discrimination at a specified score level. Such indices may be readily devised; however the mathematical formulas involved seem to be too intractable for the type of theoretical treatment appropriate here.

2. The Conditions for Maximum Discrimination

We wish to investigate the problem of how to choose test items so as to maximize the discriminating power of the test at a specified ability level. For this purpose we will assume that all items have the same value of R_i . We will then investigate what distribution of item difficulties will give optimum results under this assumption. Since D is taken to be nonnegative, we may differentiate $\log D$ with respect to g_i ($i = 1, 2, \dots, n$). Writing $N_i = N(g_i)$ we have

$$\frac{dN_i}{dg_i} = -g_i N_i, \quad (C2.1)$$

$$\frac{dP_i}{dg_i} = -N_i, \quad (C2.2)$$

$$\frac{d \log D}{dg_i} = -\frac{g_i N_i}{\sum N_i} + \frac{(1 - 2P_i)N_i}{2\sum P_i Q_i}. \quad (C2.3)$$

Setting the derivative equal to zero and transposing, we obtain n simultaneous equations:

$$\frac{(1 - 2P_i)}{g_i} = \frac{2\sum P_i Q_i}{\sum N_i} \quad (i = 1, 2, \dots, n). \quad (C2.4)$$

Since the expression on the right is the same for all n equations, we have as a condition for a maximum of D :

$$\frac{1 - 2P_1}{g_1} = \frac{1 - 2P_2}{g_2} = \dots = \frac{1 - 2P_n}{g_n}. \quad (C2.5)$$

Now the absolute value of g_1 is a single-valued function of $(1 - 2P_1)/g_1$. Consequently Equation C2.5 requires for a maximum of $\underline{D}$ that all g 's be equal in absolute value:

$$|g_1| = |g_2| = \dots = |g_n|. \quad (C2.6)$$

Under the conditions of C2.6, $P_1Q_1 = P_2Q_2 = \dots = P_nQ_n$ and $N_1 = N_2 = \dots = N_n$; so that, dropping subscripts, Equation C1.3 becomes

$$D = \frac{\sqrt{n} RN}{K \sqrt{PQ}}. \quad (C2.7)$$

Any value of g that maximizes $\underline{D}$ in Equation C2.7 will also maximize $\underline{D}$ in Equation C1.3.

It may be shown mathematically, or much more simply by numerical investigation of tabled values, that $N/\sqrt{PQ}$ is a maximum when $g = 0$. It is thus found that $\underline{D}$ is a maximum when $g_1 = 0$ for all $\underline{i}$.

When $g_1 = 0$, h_1 must equal R_1c , and $P_1 = \frac{1}{2}$. We thus reach the conclusion that if it is desired to construct a test that will have the greatest possible discriminating power for examinees at some given level of ability, $c = c_0$, then all items should be of equal difficulty such that half of those examinees whose ability score is c_0 will answer each item correctly and half will answer it incorrectly.

Similar or related conclusions have been reached by a number of writers (32, 31, 11, 1, 7, 33), starting with different premises and proceeding with varying degrees of rigor. Empirical evidence relating to this

point is also available (22, 34, 31). It should be noted, however, that the conclusion reached in the present paper may be quite different in practice from the frequent conclusion that "what is required is a set of items which average 50 per cent difficulty for the group to be tested." If, furthermore, a test is to be used to award a single scholarship to the highest-scoring examinee in a group of 100 examinees, the solution is not to use items that will be answered correctly by only one or two per cent of the examinees, but instead to use items that will be answered correctly about half the time by the two most able candidates (cf. 4, pp. 21-22).

Very few, if any, actual tests have been built in accordance with the principle deduced here.

(When the examinee has 1 chance in k of answering each item correctly by guessing, it might be expected that maximum discrimination would be found, assuming equivalent items, at the level of ability where the proportion of correct answers is $\frac{1}{2}(1 + 1/k)$. Mathematical investigation, which will not be given here, shows that this is clearly not the case. Indications are that the level of ability must be appreciably higher than this. This conclusion is plausible in view of the fact that multiple choice items become more and more unreliable as the proportion of correct answers decreases towards a chance level.)

3. Numerical Illustrations of the Discrimination Index

Figure 2, for illustrative purposes, shows the discrimination index as a function of ability for each of four hypothetical tests, together

with the regression curve of test score on ability. The discrimination indices for Tests 1, 2, and 3 were calculated from Equation C2.7. The discrimination indices for Test 4, which has a rectangular distribution of item difficulties (p_i), were found by using certain formulas published by Lawley (19). He assumes that all values of R_i are equal and that the values of h_i are normally distributed. Under these assumptions, it may be shown that the item difficulties (p_i) will be rectangularly distributed when the mean of the h 's is 0 and their standard deviation is 1 ($M_h = 0$ and $\sigma_h = 1$).

Setting $G = \sqrt{K^2 + \sigma_h^2}$ and dropping the subscript from R_i , we may write certain of Lawley's results derived from the foregoing assumptions as follows:

$$M_{s.c} = nA \left(\frac{M_h - Rc}{G} \right), \quad (C3.1)$$

and

$$\sigma_{s.c}^2 = M_{s.c} - nA_2 \left(\frac{M_h - Rc}{G}, \frac{M_h - Rc}{G}; \frac{\sigma_h^2}{G^2} \right). \quad (C3.2)$$

From C3.1 we find that

$$\frac{\partial}{\partial c} M_{s.c} = \frac{nR}{G} N \left(\frac{M_h - Rc}{G} \right). \quad (C3.3)$$

The discrimination index for a test characterized by given values of R , M_h , and σ_h may be obtained by using C3.2 and C3.3 in conjunction with the following formula (cf. C1.2):

$$D = \frac{1}{\sigma_{s.c}} \frac{\partial}{\partial c} M_{s.c}. \quad (C3.4)$$

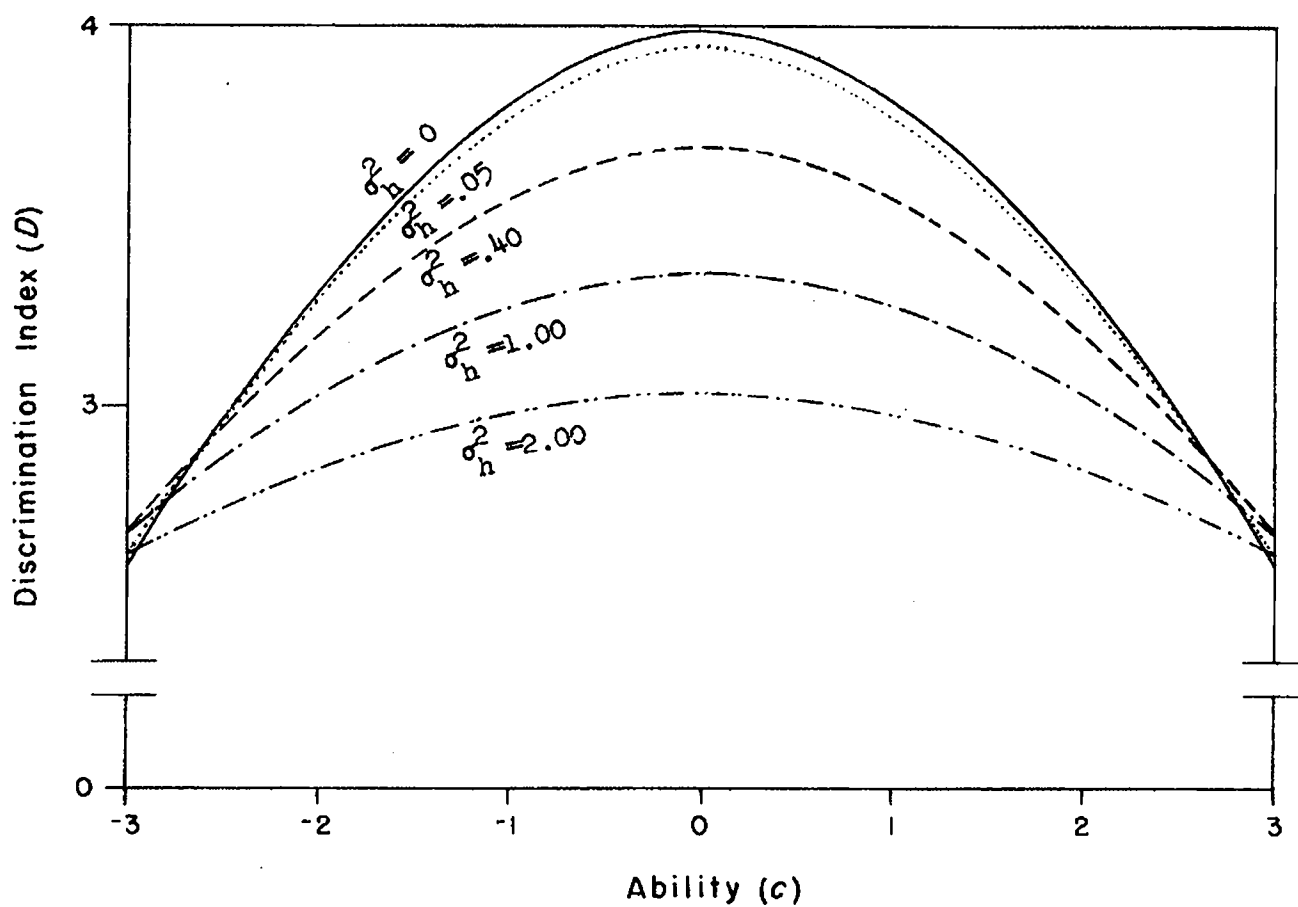


Figure 2A. The Discrimination Index as a Function of Ability, for Each of Five Tests with Specified Values of σ_h^2 .

Test 1 is a 100-item test having a reliability of only about .80 and consisting of items all of 50 per cent difficulty ($R = .243$, $\sigma_h = 0$, $M_h = 0$). The regression is almost rectilinear within the range of ability of the group tested; and the test is practically equally discriminating for all examinees within this range. Test 2 is the same as Test 1 except that its reliability is about .96 ($R = .447$; $\sigma_h = 0$, $M_h = 0$). The regression is more curved and the discriminating power of the test is much higher for examinees of average ability than for examinees at the extremes of ability in the group tested. The standard error of measurement at different levels of ability is indicated by a dotted line for this test. Test 3 may be considered to be the same test as Test 2, but administered to a less competent group of examinees ($R = .447$, $\sigma_h = 0$, $M_h = 1.07$). We see here what happens when a test is too difficult for the group tested: the regression of test score on ability is distinctly non-linear, and the test has low discriminating power for the less competent examinees. The curves for Test 3 are actually the same as, or continuations of, the curves for Test 2, except that the scale on the base line is changed. Test 4 is the same as Test 2 except that the items here are not all of the same difficulty, the item difficulties (p_i) in Test 4 being rectangularly distributed ($R = .447$, $\sigma_h = 1$, $M_h = 0$).

For purposes of comparison, Figure 2A superimposes the curves showing the discrimination index for each of five tests. The test for which $\sigma_h^2 = 0$ is the same as Test 2 in Figure 2. The values of h_i in each of the other tests are normally distributed with $M_h = 0$, and the value of R_i for each item is in all cases .447. Each of the five tests consists of 100 items.

The tests differ from each other only in the variability of the difficulties of the test items, as measured by σ_h^2 . The test for which $\sigma_h^2 = 1.00$ is the same as Test 4 in Figure 2 -- the values of p_1 in this test have a rectangular distribution. The test for which $\sigma_h^2 = .40$ is the one most nearly like the usual type of test having a moderate spread of item difficulty.

Figure 2A shows that the test composed solely of items of 50 per cent difficulty is more discriminating than any of the other tests for examinees at any level of ability between $c = -2.5$ and $c = +2.5$. Since only about one per cent of the examinees will lie outside this range of ability, these results suggest that many of our tests could be improved by restricting the range of item difficulty. (The reader may wish to refer to Davis (4) and Flanagan (9) for arguments opposing this point of view.)

4. Relation of the Discrimination Index to the Test Reliability

Let us consider the following function of D and expand it by means of Equation C3.4, using the symbol b_{sc} for the derivative of $M_{s.c}$ with respect to c :

$$\frac{D^2}{1 + D^2} = \frac{b_{sc}^2}{\sigma_{s.c}^2 + b_{sc}^2} . \quad (C4.1)$$

Suppose now that the range of ability present in some special group of examinees tested is sufficiently narrow so that within this range the values of b_{sc} and of $\sigma_{s.c}$ may be treated as constants for some hypothetical test. In this case b_{sc} is the regression coefficient in the equation for predicting s from c , and consequently, by the usual formula for a regression coefficient:

$$b_{sc} = \frac{\sigma_s}{\sigma_c} r_{sc} . \quad (C4.2)$$

Similarly, $\sigma_{s.c}$ is the standard error of estimate, so we may write

$$\sigma_{s.c}^2 = \sigma_s^2 (1 - r_{sc}^2) . \quad (C4.3)$$

If we choose the units of measurement for c so that $\sigma_c = 1$ and then substitute C4.2 and C4.3 in C4.1, we obtain for this special case

$$\frac{D^2}{1 + D^2} = r_{sc}^2 . \quad (C4.4)$$

When the regression of s on c is approximately linear, $r_{sc}^2 = r_{st}^2 = r_{ss}$ (see Section B7); so finally, for this special case,

$$\frac{D^2}{1 + D^2} = r_{ss} . \quad (C4.5)$$

This result suggests that the function $D^2/(1 + D^2)$ might well be used as a measure of the discriminating power of the test at a given level of ability. The interpretation of this function would be facilitated by its equivalence to the reliability coefficient of a certain hypothetical test.

D. The Frequency Distribution of Test Scores

The univariate distribution of test scores, f_s , may be obtained by integrating Equation B2.2 according to the usual procedure:

$$f_s = \int_{-\infty}^{\infty} f_{cs} dc = \Sigma^* \int_{-\infty}^{\infty} f_c \prod_s P_i \prod_{n-s} Q_i dc . \quad (D1.1)$$

Another special case that is of interest occurs when all items are of fifty per cent difficulty ($h = 0$) and all values of R_i are equal to $1/\sqrt{2}$. In this case $g_i = -c$, and D1.2 becomes

$$f_s = \binom{n}{s} \int_{-\infty}^{\infty} N(c) A^s(-c) B^{n-s}(-c) dc, \quad (D1.4)$$

where the superscripts are exponents and $B(-c) = 1 - A(-c)$. Now

$$\frac{d}{dc} A(-c) = N(c); \quad (D1.5)$$

also $A(-c) = 1$ when $c = \infty$ and $A(-c) = 0$ when $c = -\infty$. Consequently, writing $A = A(-c)$ and replacing $N(c)dc$ in D1.4 by dA , we obtain the result

$$f_s = \binom{n}{s} \int_0^1 A^s (1 - A)^{n-s} dA. \quad (D1.6)$$

The integral in D1.6 is a beta function; so finally,

$$f_s = \frac{n!}{s!(n-s)!} \frac{s!(n-s)!}{(n+1)!} = \frac{1}{n+1}. \quad (D1.7)$$

Since s can only vary from 0 to n , we have the result that in this special case s is rectangularly distributed.

The value for many general purposes of having an approximately rectangular distribution of test scores has been pointed out by Ferguson (7). In a previous article (14), Jackson and Ferguson emphasize the value of having different specified shapes of score distributions for different specified purposes. For example, if the test is to be used merely to separate examinees into a successful group and a failing group, it is

desirable to have as few examinees as possible with scores near the cutting score. The most desirable distribution of scores for this purpose would therefore be a U-shaped distribution with the antimode at the cutting score. Equation D1.2 will give such U-shaped distributions whenever the item intercorrelations are sufficiently high. Apparently, when c is normally distributed, R_1 must be greater than $1/\sqrt{2}$ before this can occur, however. Such large values of R_1 are seldom obtained with most types of cognitive tests.

The results given are sufficient to show that the distribution of test scores cannot in general be expected to be normal, or even approximately normal. The question naturally arises as to what possible shapes the frequency distribution f_s , as given in Equation D1.2, may assume. The answer is that this function may assume any shape whatsoever, provided the item intercorrelations are sufficiently high.

Suppose, for example, that it is desired to obtain the following arbitrarily chosen frequency distribution for a five-item test:

<u>s</u>	<u>f_s</u>	<u>cumulative frequency</u>
5	.25	1.00
4	0	.75
3	.25	.75
2	.32	.50
1	.16	.18
0	.02	.02
	<u>1.00</u>	

Referring to a table of the normal curve we find that the required cumulative frequencies correspond to relative deviates of ∞ , +0.67, +0.67, +0.00, -0.92, and -2.05. A little thought will show that if we select five items

that are perfectly correlated with $\underline{c}$ and that have values of h_i equal to the last five of the six relative deviates just listed, the resulting five-item test will have the required frequency distribution of test scores. The same result may be obtained from Equation D1.2. This result, although of course useless from a practical point of view, is given in order to show that there is really no limit to the different shapes that may be assumed by the frequency function of Equation D1.2, provided the item intercorrelations are sufficiently high. When the item intercorrelations are low, however, the test score distribution is necessarily not very different from the generalized binomial (D1.3), i.e., it is necessarily bell-shaped.

Next let us investigate the moments of f_s . The mean of the test score distribution can be found by simple algebra without making any assumptions. As is well known, the mean test score is $\underline{n}$ times the mean item difficulty:

$$M_s = \sum p_i . \quad (D1.8)$$

This same result may be derived from Equation D1.2, as follows:

$$\begin{aligned} M_s &= \sum_{s=0}^n s f_s = \int_{-\infty}^{\infty} N(c) \sum_s s \prod_{i=1}^s P_i \prod_{i=s+1}^n Q_i dc \\ &= \int_{-\infty}^{\infty} N(c) M_{s.c} dc . \end{aligned} \quad (D1.9)$$

Substituting the value for $M_{s.c}$ from B3.1, we have

$$M_s = \sum_i \int_{-\infty}^{\infty} N(c) P_i dc . \quad (D1.10)$$

This integral may be evaluated with a result that agrees with that obtained algebraically and presented in D1.8.

The standard deviation of the test score distribution may likewise be derived without assumptions by simple algebra, as shown in Equation B6.4. This same result also may be obtained from Equation D1.2 by evaluation of the integrals in the formula

$$\sigma_s^2 = \sum_{s=0}^n s^2 f_s - M_s^2. \quad (D1.11)$$

The third and higher order moments of the test score distribution cannot be expressed as simple functions of the item difficulties and inter-correlations. These moments can be obtained by numerical integration, or by expanding the necessary integrals by means of the infinite series developed by Pearson (27), but the result is too cumbersome to be of much practical value for present purposes.

In view of the obstacles encountered here, it appears that the best practical method of obtaining further insight into the general relation between the item statistics and the shape of the distribution of test scores is to examine the limiting frequency distribution approached by the score distribution as the number of items becomes very large. This will be done in the following section.

It is of interest, now that we have obtained a formula for the frequency distribution of test scores, to note that the line of reasoning followed for this purpose is, in broad outline, the same as that used by Lazarsfeld (21) in dealing with the problem of latent structure, but

opposite in direction. We have made certain assumptions as to the shape of the item characteristic curve and have derived the distribution of test scores, on the basis of these assumptions, as a function of the frequency distribution of the underlying ability in the group of examinees tested. Lazarsfeld, on the other hand, starts with certain assumptions as to the shape of the item characteristic curve ("trace line") and attempts to derive the distribution of the underlying ability (trait) in the group of examinees tested as a function of the known distribution of actual test scores. No entirely satisfactory general solution to the problem of latent structure has as yet been found.

E. The Limiting Frequency Distribution of Test Scores for Large n --

The Frequency Distribution of True Scores

1. Derivation of the Distribution

In discussing the limiting distribution of scores as n becomes very large, we will concern ourselves only with the relative score, $z = s/n$. When n is infinite, z becomes the same as the relative true score, which is denoted by t .

We saw from Equation B5.5 that as n becomes very large, all the frequency of the bivariate distribution of z and c becomes concentrated at the regression lines, which in this case coincide. Thus for infinite n ,

$$t = M_{z.c} . \quad (E1.1)$$

In the case of relative scores, the regression is seen from B3.1 to be given by the average value of $\underline{P}$:

$$M_{z.c} = \frac{1}{n} \sum P_1 = M_P . \quad (E1.2)$$

Let us replace M_P by the expression $M(c)$ to emphasize the fact that M_P is being considered as a function of c . We see from E1.1 and E1.2 that the distribution of $\underline{t}$ is the same as the distribution of the function $M(c)$. If we assume that $R_1 \neq 0$, we may denote the inverse function of $M(c)$ by $M^{-1}(t)$, so that $c = M^{-1}(t)$. If c has the frequency distribution $f(c)$, the distribution of $t = M(c)$ may be found by the usual methods to be

$$f_t = f \left[M^{-1}(t) \right] \frac{\partial M^{-1}(t)}{\partial t} . \quad (E1.3)$$

When c is normally distributed, we have

$$f_t = N \left[M^{-1}(t) \right] \frac{\partial M^{-1}(t)}{\partial t} . \quad (E1.4)$$

Let us suppose that an infinite number of equivalent $\underline{m}$ -item tests are administered, so that $\underline{t}$ is the average relative score obtained on all these tests. Since the tests are equivalent we have, as before,

$$t = M(c) = \frac{1}{m} \sum_{i=1}^m P_i . \quad (E1.5)$$

We find that (cf. Equation C1.3)

$$\frac{\partial t}{\partial c} = \frac{1}{m} \sum \frac{R_i N(g_i)}{K_i} . \quad (E1.6)$$

Since $c = M^{-1}(t)$, $\frac{\partial M^{-1}(t)}{\partial t}$ is the reciprocal of the expression on the right

side of Equation E1.6. Substituting this result in E1.4, we obtain finally:

$$f_t = \frac{mN[M^{-1}(t)]}{\sum \frac{m R_1 N(g_1)}{K_1}} . \quad (\text{E1.7})$$

In E1.7, g_1 is as usual (A3.5) a function of h_1 , R_1 , and c ; but here c is treated as a function of t , so that g_1 is a function of t . For purposes of calculation, f_t may perhaps best be computed from the parametric equations:

$$f_t = \frac{mN(c)}{\sum \frac{m R_1 N(g_1)}{K_1}} , \quad t = M(c) , \quad (\text{E1.8})$$

where c is the parameter. When the values of h_1 and R_1 are specified, actual values of the frequency distribution of relative true scores can be calculated from these equations. This has been done for certain simple illustrative examples in the following section.

The cumulative frequency distribution of t , which we will denote by $F(t)$, may be of interest; it is found from E1.4, by the usual methods, to be

$$F(t) = \int_{-\infty}^t f_t dt = \int_{-\infty}^{M^{-1}(t)} N(c) dc . \quad (\text{E1.9})$$

The second integral is obtained from the first by the substitution $c = M^{-1}(t)$. Actual values of the cumulative frequency function of relative

true scores can be obtained fairly easily from E1.9.

The frequency distribution of relative scores actually approaches the frequency distribution f_t as a limit when $\underline{n}$ becomes infinitely large. This same result may be obtained from Equation D1.2 by finding the limit of the characteristic function as $\underline{n}$ becomes infinite. Unfortunately, it has not been possible to date to obtain any simple expression for the amount of the discrepancy between f_z , for any given $\underline{n}$, and f_t .

We may note at this point that, because of the relation between $\underline{t}$ and $\underline{c}$, the distribution ($f_{s.t}$) of actual scores for a fixed true score is the same as the distribution ($f_{s.c}$) of actual scores for the corresponding fixed ability score. Consequently, $f_{s.t}$ is the same generalized binomial as is given by Equation B1.2.

2. Illustrative Examples of the Distribution of Relative True Scores

Relative true scores are the relative scores that would be obtained if the test were a perfectly reliable measuring instrument. The algebraic difference between an examinee's actual score and his true score is the error of measurement. As indicated in the preceding paragraph, the errors of measurement for any given true score have a generalized binomial distribution. The errors of measurement are not independent of the true score, since their standard deviation and their frequency distribution are different for different true scores. The usual proof that the errors of measurement are uncorrelated with true scores remains valid in the present context, however.

Figure 3 illustrates the effect on the shape of the true score dis-

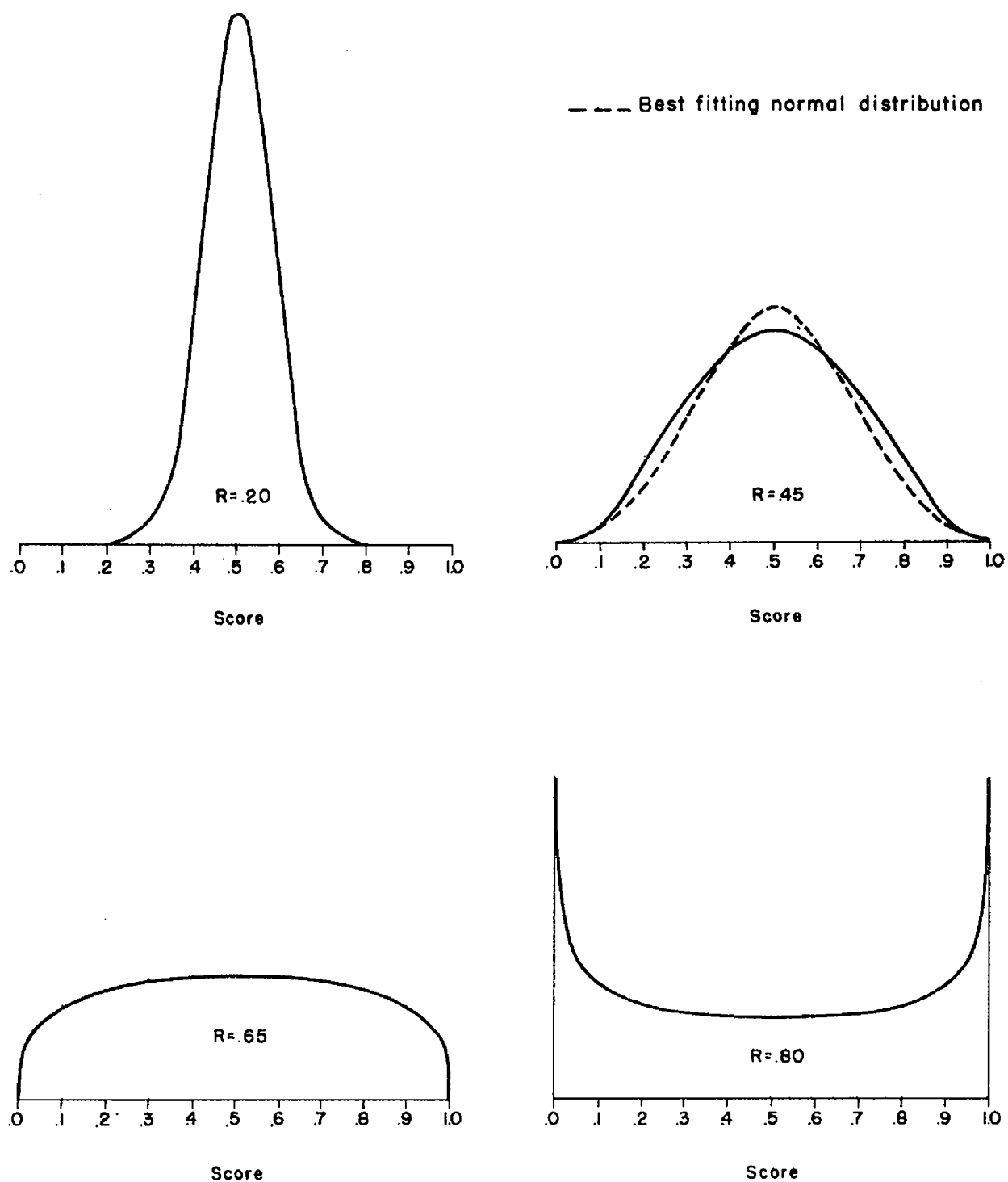


Figure 3. Frequency Distributions of Relative Scores on Four Infinitely Long Tests Composed of Equivalent Items of 50 Per Cent Difficulty Whose Correlations with Ability Are as Indicated.

tribution of increasing the correlation (R) of the items with ability. The values of f_t needed for drawing these distributions were obtained from Equation E1.7. The distributions shown are for relative true scores on tests composed of equivalent items of 50 per cent difficulty. When $R = .20$, the distribution, although platykurtic, is so nearly normal that the best fitting normal distribution can hardly be drawn on the same graph; the standard deviation is low, and only about a half of the possible range of scores is found to occur with any frequency in actual practice. When $R = .45$, the distribution of scores is still nearly normal, and the entire range of scores is utilized. When $R = .65$, the standard deviation has increased further and the distribution, although still unimodal, is nearly rectangular. The distribution becomes actually rectangular when $R = .707$, as was shown by Equation D1.7. As the correlation increases above .707, the distribution becomes more and more U-shaped, as indicated by the distribution for $R = .80$.

Actual tests having an average item difficulty of about 50 per cent would produce frequency distributions similar to the true score distributions illustrated. If the average item difficulty differed appreciably from 50 per cent, the actual score distributions would still resemble those illustrated except that they would not be symmetric.

F. The Bivariate Distribution of Scores on Two Tests Measuring the Same Ability

If the items in two tests have only a single common factor, the distributions of the test scores (s_1 and s_2) on the two tests will be inde-

pendent when $\underline{c}$ is fixed. The conditional bivariate distribution of test scores will therefore be

$$f_{s_1 s_2 \cdot c} = f_{s_1 \cdot c} f_{s_2 \cdot c} . \quad (F1.1)$$

The right side of this equation may be evaluated by B1.2.

The trivariate distribution of s_1 , s_2 , and $\underline{c}$ is therefore

$$f_{cs_1 s_2} = N(c) f_{s_1 s_2 \cdot c} ; \quad (F1.2)$$

and the bivariate (marginal) distribution of the two test scores can be found from this by integrating out the variable $\underline{c}$:

$$f_{s_1 s_2} = \int_{-\infty}^{\infty} N(c) f_{s_1 s_2 \cdot c} dc . \quad (F1.3)$$

This result is too cumbersome in expanded notation for us to draw any general theoretical conclusions from it. For a given set of data, however, equation F1.3 can be handled by means of numerical integration. This has actually been done as part of the empirical investigation reported in the following sections. The resulting bivariate score distributions are given in Tables 24 to 30.

In the special case where all items are of 50 per cent difficulty and $R_1 = \sqrt{.5}$ for all items, the bivariate distribution of test scores is found by direct integration (cf. D1.7) to be

$$f_{s_1 s_2} = \frac{1}{n_1 + n_2 + 1} \frac{\binom{n_1}{s_1} \binom{n_2}{s_2}}{\binom{n_1 + n_2}{s_1 + s_2}} . \quad (\text{Fl.4})$$

(It may be noted that the second fraction in Fl.4 is the general term of a hypergeometric distribution.) The bivariate distribution of Fl.4 is somewhat of a curiosity. The marginal distributions of s_1 and s_2 are of course rectangular (Dl.7). The distribution of $s_1 + s_2$ (obtained by summing the frequencies in Fl.4 diagonally) is also rectangular. The bivariate distribution has two modes: one at the corner where $s_1 = s_2 = 0$; the other at the corner where $s_1 = n_1$, $s_2 = n_2$. The two antimodes are at the two remaining corners. The frequency surface represented by Fl.4 may be well characterized as saddle-shaped.

The regression of s_2 on s_1 for the special frequency surface of Fl.4 is found to be

$$M_{s_2 \cdot s_1} = \frac{n_2(s_1 + 1)}{n_1 + 2} . \quad (\text{Fl.5})$$

It is obvious that in this special case the regression is linear. Examination of the appropriate formula shows, however, that in general the regression of one test score on another will not be strictly linear, even though the two tests are parallel forms. This conclusion is confirmed by the numerical integration of the appropriate formula for hypothetical values of h_i and R_i chosen to represent two strictly parallel tests.

G. Extension of Results to Multiple Choice Items

When it is possible for the examinee to obtain the correct answer to a test item by sheer guessing, it is no longer reasonable to assume that the item characteristic curve is a normal ogive. No matter how low the ability of an examinee may be, he still has some appreciable chance of answering such items correctly. The formulas derived in the present article in general cannot be applied to multiple choice tests, or to other tests where guessing plays a significant role. Formulas parallel to those given here have been worked out for the multiple choice case on the assumption that guessing, whenever it occurs at all, is purely random. Such an assumption is certainly only a rough approximation to the real life situation. No empirical check has as yet been made to determine to what extent the formulas derived on the basis of this assumption yield adequate predictions of observed results. The formulas themselves will not be presented here.

III. EMPIRICAL VERIFICATION

H. The Plan

The theoretical results of Part II, Sections B and C, relating test score to ability, cannot be checked empirically except by the use of a test so long and so reliable that scores on this test, after being normalized or transformed in some other way, can be satisfactorily substituted for the unknown values of the ability score, $\underline{c}$. In view of the difficulty of securing a sufficiently long test that meets the other criteria desirable for a first empirical study, the empirical work done to date has related only to the theoretical results of Part II, Sections D and F, where no measure of ability is involved in the final results. The results to be checked are those of Equations D1.2 and F1.3, which give the univariate frequency distribution of the test score and the bivariate distribution of two scores.

The empirical procedure involves essentially: 1) selection of a number of short tests for which the examinees' answer sheets are available; 2) computation of the item statistics p_i and R_i ; 3) computation of f_s and of $f_{s_1 s_2}$ by means of Equations D1.2 and F1.3; and 4) comparison of the theoretical results so obtained with the actual univariate and bivariate distributions of test scores.

I. The Data

In selecting test data for the empirical study, three considerations were given paramount importance: 1) the test must be composed of free response items such that guessing by the examinee could have little effect

on the correctness of his responses; 2) the test items must in so far as possible have only one common factor; and 3) the test must have been administered to a group of examinees of such a nature that it would not be unreasonable to assume ability (c) to be normally distributed in the group tested. These three considerations limited the choice among the readily available test data so drastically that other considerations were largely disregarded.

It may be well to call attention to the third consideration. It is obviously impossible to predict anything about the shape of the distribution of scores for a given group of examinees if nothing is known or assumed about the nature of the group tested, since the distribution of ability, and hence the distribution of scores, may take any shape whatsoever. It would be possible and desirable to use the method of Section A2 to determine the actual frequency distribution of ability in the group of examinees; however, this method would have required the use of a very long test. In the absence of such a test, it was necessary to assume that ability, as defined by the variable c , is at least approximately normally distributed in the group tested. The success of the empirical verification undertaken here may depend to a considerable extent on the appropriateness of this assumption.

The data used for the empirical verification were kindly provided by Dr. Lynnette B. Plumlee, who had carefully collected them for other research purposes (30). The available data were the responses of each of four groups of male examinees to 80 free response mathematics items covering the fields of algebra, plane geometry, and trigonometry as taught in

the usual high school. The 80 items had been originally chosen so as to represent the entire range of difficulty from $p_i = .10$ to $p_i = .90$.

Half of the 80 items were discarded since time apparently did not permit all examinees to attempt to answer them. The 40 remaining items were next classified on the basis of subjective judgment into the following categories: 1) verbally stated problems to be solved by algebra, 2) algebra items stated in mathematical form, 3) geometry items, and 4) trigonometry items. In order to minimize the presence of group factors common to some items but not to all, all except one of the items in each of the first, third, and fourth categories were discarded, leaving 28 items for empirical study, 25 of which were algebra items stated in mathematical form. The 25 algebra items covered elementary, intermediate, and, to some extent, advanced algebra. The 3 remaining items represented three different areas, and hence seemed unlikely to introduce any additional group factors.

Since the items had been administered to each of four groups of examinees, it was desired to select one group for further study. For this purpose the matrix of tetrachoric item intercorrelations was computed for the first five test items, separately for each of the four groups of examinees. An attempt was made to determine the communalities, separately for each group, under the assumption of a single common factor. The method of triads and the iterated centroid method were used. In three of the four groups one or more of the communalities were found either to exceed 1.0 or to be very close to 1.0. This result could be attributed to any one of the following causes: 1) the very large sampling error of some

of the tetrachorics; 2) the existence of more than one common factor among the items; 3) non-normality of the distribution of ability in the group tested, making the use of the tetrachoric correlation coefficient inappropriate.

The fourth matrix yielded reasonable communalities. On this basis it was decided to use the fourth group, composed of 136 examinees, for all further empirical work.

J. Procedure

1. Calculating the Item Statistics

The matrix of tetrachoric intercorrelations for the twenty-eight items, as calculated for the fourth group, is presented in Table 4. No entry was made in the correlation matrix for nine cases where there was zero frequency in one of the cells of the fourfold table used to calculate the tetrachoric correlation. (Strictly speaking, the tetrachoric correlation is ± 1.00 for these nine cases -- -1.00 in the case of the correlation between items 54 and 6, $+1.00$ in each of the other eight cases. The sampling error of these values is obviously excessive.)

The common factor loadings of the items were obtained by the following procedure: 1) Estimated communalities were entered in the diagonal. 2) The square root of the product of the appropriate estimated communalities was entered in each of the nine empty cells of the matrix. 3) Factor loadings on a single factor were obtained by the centroid method. 4) The square of the factor loading of each item was used as a new estimate of its communality, and the whole process was iterated until the factor loadings

remained unchanged. The resulting factor loadings, together with the item difficulties, are given in Table 5. These two values for each item are the only item statistics required by the formulas that have been developed in the preceding sections.

Table 6 gives the residual correlations after extraction of the first factor. The standard deviation (computed from a grouped frequency distribution) of all 378 residuals, including the 9 listed in the footnote to the table, was found to be .17. This value may be compared to the sampling error of a tetrachoric correlation when the sample size is 136 and the true correlation is .30. Table 1 gives a few values of this sampling error for various combinations of item difficulties, as computed from the table provided by Hayes (13).

In view of these standard errors, the obtained standard deviation of the residuals indicates that any group factors in the matrix are of minor relative importance. This conclusion is borne out by a detailed study of the residuals, and also by the extraction of a second centroid factor after inserting estimated diagonal entries in the residual matrix. The sum of squares of the unrotated second factor loadings was found to be only 1.6 as compared to the corresponding value of 7.5 for the first factor loadings, indicating that the contribution of the second factor was relatively small.

It is unfortunate that the residuals are as large as they are, since this result means that the matrix of tetrachoric intercorrelations cannot be accurately determined from the values of R_i , as required by the deriva-

Table 1

Standard Error of the Tetrachoric Correlation
between Two Items, for Specified Combinations
of Item Difficulties, When the True Correlation
Is .30 and the Number of Cases Is 136.

<u>Item Difficulty</u>	<u>Item Difficulty</u>		
	<u>.16</u>	<u>.50</u>	<u>.84</u>
.16	.20	.15	.17
.50	.15	.13	.15
.84	.17	.15	.20

tions in the preceding sections. This difficulty could presumably have been avoided if more cases had been available from which to compute the tetrachoric correlations.

2. Selection of Tests

Eight different combinations of about ten items each were selected to serve as short tests to be used in the empirical verification. The allocation of the items to each of the eight tests is indicated in Table 7, together with the difficulties and common factor loadings of the items. The eight tests may be briefly summarized as follows:

Test 2 -- difficult items ($p_1 = .09$ to $p_1 = .34$).

Test 5 -- items of medium difficulty ($p_1 = .43$ to $p_1 = .68$).

Test 8 -- easy items ($p_1 = .71$ to $p_1 = .90$).

Test 8' -- like Test 8, but slightly shorter.

Test 82 -- five "easy" and five "difficult" items.

Test h -- highly discriminating items, irrespective of difficulty
($R_1 = .55$ or more).

Test L -- poorly discriminating items, irrespective of difficulty
($R_1 = .47$ or less).

Test r -- item difficulties rectangularly distributed.

The bivariate frequency distributions of test scores were studied for the following seven pairs of tests: 2 and 5, 2 and 8, 5 and 8, 5 and 82, 8' and r, 82 and r, h and L. No bivariate distribution between any two tests containing identical items was investigated.

3. Obtaining the Theoretical Bivariate Frequency Distribution of Test Score and Ability

Fifteen values of $\underline{c}$ were selected for further calculations: $c = -3.5, -3.0, -2.5, -2.0, -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5$. For each item the ordinate (P_i) of the item characteristic curve was obtained for each of these values of $\underline{c}$ from Equations A3.3, A3.4, A3.5 and A3.6.

For each test, for each value of $\underline{c}$, the theoretical conditional distribution of test scores was calculated from B1.2. An example for $n = 3$ will clarify the method used. Suppose (for a given value of $\underline{c}$) $P_1 = .5, P_2 = .4, P_3 = .2$. We know that the desired conditional distribution is given by the appropriate successive terms of the expansion of the product $(.5 + .5)(.4 + .6)(.2 + .8)$. Expanding this product while keeping separate the terms corresponding to the separate values of $s = 0, 1, 2, 3$, we have

$$\begin{aligned} (.5 + .5)(.4 + .6)(.2 + .8) &= (.2 + .5 + .3)(.2 + .8) \\ &= .04 + .26 + .46 + .24 . \end{aligned}$$

The four terms obtained in this way are the conditional frequencies of occurrence for $s = 3, 2, 1$, and 0 respectively. The conditional probability that all three items will be answered correctly, for example, is given by the first term, .04.

It is convenient to consider the $n + 1$ conditional frequencies of $\underline{s}$ for each of the 15 values of $\underline{c}$ for any given test as forming a matrix, $F_{s.c}$, with $n + 1$ rows and 15 columns. If D_v is a diagonal matrix whose elements are the fifteen values of $N(c)$, the matrix F_{sc} of the ordinates

of the bivariate frequency distribution of $\underline{s}$ and $\underline{c}$ is seen from Equation B2.1 to be

$$F_{sc} = F_{s \cdot c} D_v .$$

Tables 8 through 15 present, for the selected values of $\underline{c}$, the theoretical ordinates of the bivariate distribution of test score and ability for each of the eight tests. It should be noted that the values given in the table are ordinates for the stated values of $\underline{c}$, not cell frequencies corresponding to class intervals of $\underline{c}$; consequently, the sum of a row of tabled values does not give the marginal frequency of the test score.

4. Obtaining the Theoretical Univariate Distributions of Test Scores

The theoretical univariate distribution of each test score may be obtained from the bivariate distribution of test score and ability by integration, as indicated in Equation D1.2. It was found most satisfactory in the present case to perform the numerical integration by use of the trapezoidal rule:

$$\int_{c_o}^{c_m} f dc = h \left(\frac{f_o}{2} + f_1 + f_2 + \dots + \frac{f_m}{2} \right) ,$$

where $\underline{f}$ is the function of $\underline{c}$ to be integrated, f_u is the value of $\underline{f}$ when $c = c_u$, c_m is the upper limit of integration, c_o is the lower limit of integration, and h is the difference between any two successive values of $\underline{c}$, c_u and c_{u+1} . In the present case, c_o and c_m may be taken as -4.0 and $+4.0$, respectively, so that the ordinates f_o and f_m are so small that they

may be ignored. Since in our case $h = \frac{1}{2}$, the trapezoidal rule for our purposes therefore reduces to

$$\int_{-\infty}^{\infty} f dc = \frac{1}{2} \sum_{c_u=-3.5}^{c_u=3.5} f_u .$$

The adequacy of this method of numerical integration for present purposes was carefully checked.

The univariate frequency distribution of test score may thus be obtained by summing separately each row of F_{sc} and dividing by 2. These distributions are shown in the right hand columns of Tables 8 through 15. The comparison of these theoretical distributions with the corresponding actual distributions will be made at a later point.

5. Obtaining the Theoretical Bivariate Frequency Distribution for Two Subtest Scores

The theoretical bivariate distribution for two subtest scores may be obtained from Equation F1.3. This distribution may be considered to be represented by a matrix $F_{s_1 s_2}$ having $n_1 + 1$ rows and $n_2 + 1$ columns. If the integration of Equation F1.3 is carried out by means of the trapezoidal rule, we have

$$F_{s_1 s_2} = \frac{1}{2} F_{s_1 c} F'_{s_2 \cdot c} = \frac{1}{2} F_{s_1 \cdot c} F'_{s_2 c} ,$$

where the primed matrices are transposed.

The theoretical distributions of test scores obtained in this way are presented and discussed in Section K2.

L. Comparison of Theoretical and Actual Results

1. Comparison of Univariate Frequency Distributions of Test Score

The theoretical and obtained univariate distributions of scores are compared graphically in Figures 4 through 11, and numerically in Tables 16 through 23. The tables also show the means and standard deviations of the distributions, together with a measure of skewness,

$$\alpha_3 = \frac{1}{N\sigma_s^3} \sum (s - M_s)^3 ,$$

where $N = 136$; and a measure of kurtosis,

$$\beta_2 - 3 = \frac{1}{N\sigma_s^4} \sum (s - M_s)^4 - 3 .$$

For a symmetrical curve, $\alpha_3 = 0$; positive values of α_3 result from positive skewness, negative values from negative skewness. For a normal curve, $\beta_2 - 3 = 0$; positive values of $\beta_2 - 3$ result from leptokurtosis, negative values from platykurtosis.

We may pause here for a moment in order to note a few facts about the shapes of the obtained score distributions: 1) the test composed of difficult items produces a positively skewed distribution of scores; the tests composed of easy items produce negatively skewed distributions. 2) Leaving these three highly skewed distributions out of consideration, we see that Test 82 has a mesokurtic distribution and that all other distributions are at least somewhat platykurtic; that the highly discriminating items produce a much more platykurtic distribution than do the

poorly discriminating items; and that the most platykurtic distribution is produced by Test 5, which is composed entirely of items of 50 per cent difficulty. 3) Although the tests are not all of exactly the same length, it is obvious that the tests that contain many very easy or very difficult items yield scores with relatively small standard deviations, whereas Test 5, composed entirely of 50 per cent difficulty items, produces scores having a relatively large standard deviation; also there is a particularly striking difference between Test h (composed of 10 highly discriminating items), which yields scores having a standard deviation of 2.44, and Test L (composed of 10 poorly discriminating items), which yields scores having a standard deviation of 1.59. These results are specific to the tests at hand, but the conclusions reached appear to be capable of considerable generalization.

The standard error of α_3 for a sample of 136 cases drawn at random from a normal population is 0.21; the standard error of $A_2 - 3$ in the same situation 0.41. Several of the distributions of test scores presented here are distinctly nonnormal, however, and these standard errors cannot be considered to be applicable in these cases. It should be noted, furthermore, that no test of the statistical significance of the difference between theoretical and observed statistics can be made in the present situation, in view of the fact that the theoretical values have been calculated from the observed data in such a way as to obscure completely the number of degrees of freedom remaining for any statistical significance test. It is not even possible to set a lower limit (above zero) to the number of degrees of freedom remaining.

This situation is actually not a disadvantage, since we are really not concerned with the statistical significance of the difference between theoretical and observed distributions. It might as well be admitted from the start that the assumptions underlying the theory will never be completely fulfilled by any set of data; and that consequently we are bound to find statistically significant differences if only enough cases are used. What we really are concerned with is whether or not the difference between theoretical and observed distributions is of practical significance. Since this question cannot be reduced to a statistical test, the reader can perhaps best satisfy himself by a visual examination of the figures presented.

The standard errors quoted, however, can be used to throw some light on the practical significance of the differences in question as long as we bear in mind that no significance test is being made. We may, for example, note that scores on Test 82 (see Table 20) are approximately normally distributed, and that the theoretical and observed values of α_3 differ by an amount equal to 0.4 times the standard error (for random samples of 136 cases drawn from a normal population) of the observed value. Now α_3 is normally distributed (in such random samples), and only 30 per cent of the cases in a normal distribution fall between -0.4 and +0.4 times the standard deviation. Consequently, we may conclude that in 70 samples out of 100 the observed sample value of α_3 will differ from the true value in the normal population from which the samples were drawn by an amount greater than the obtained difference between the observed and the theoretical values. This statement provides a basis of

comparison for estimating the practical significance of the obtained difference. Since two samples, on the average, will differ from each other to a greater extent than one sample will differ from the population from which it is drawn, we may state in this case, for example, that we may expect the theoretical prediction of the value of α_3 to be considerably more accurate than would be a prediction based on the value of α_3 actually observed in a previously drawn sample of 136 cases.

We may note that in no case do the theoretical values of α_3 differ from the observed values by an amount as large as the standard error of the observed values in samples from a normal population. The same statement may be made for the standard deviation of the test scores (σ); and also for β_2 , except in the case of Test 8', where the difference is slightly greater than the standard error. In the case of the mean (M), the theoretical and the observed values are in every case identical within a margin of .01.

Chi-square values have been calculated for the difference between theoretical and observed distributions, as shown in Tables 16 through 23. In calculating these values, enough class intervals at the extremes of the distributions were combined to secure a theoretical frequency of at least 5 for each interval. It must be emphasized that these values do not provide tests of statistical significance, since the degrees of freedom appropriate for this purpose are unknown. The values of chi-square can be used, however, to provide a useful comparison for evaluating the practical significance of the difference.

In the case of Test 2, for example, a value of 3.62 was obtained for chi-square from a grouped score distribution having six class intervals. Entering a table of chi-square values, we find that in random sampling chi-square with 5 degrees of freedom will exceed 3.62 about 61 per cent of the time. Thus we may say that the obtained difference between theoretical and observed score distributions for Test 2 is no greater than the difference that would be found between population and sample distributions in about 61 per cent of the samples. A fortiori, in this case, the theoretical distribution provides a better prediction of the observed distribution than could be obtained on the average from the observed score distribution of a second random sample of 136 examinees drawn from the same population of examinees.

The probabilities that the values of chi-square obtained in Tables 16 through 23 would be exceeded in random sampling are given in Table 2 for the case where the population parameters are not estimated from the data. (These probabilities are higher than would be appropriate if the difference between theoretical and observed frequency distributions were being tested for significance.)

The only test that has a rather large value of chi-square is Test h. The theoretical and observed distributions of scores on this test have means, standard deviations, and values of α_3 and β_2 that are very much alike, however. Examination of the histograms of these two distributions (Figure 9) seems to indicate that they differ only in an unsystematic fashion. In fact the theoretical distribution seems to provide about as good a fit to the observed distribution as could be desired in view of the irregularities of the latter.

Table 2

Probability that Certain Values of Chi-Square, Having the Specified Degrees of Freedom, Will Be Exceeded in Random Sampling.

<u>Chi-Square</u>	<u>Degrees of Freedom</u>	<u>Probability</u>
3.6 (Test 2)	5	.61
5.5 (Test 5)	7	.60
.8 (Test 8)	6	.99
.7 (Test 8')	6	.97
3.8 (Test 82)	6	.70
18.4 (Test h)	9	.03
2.0 (Test L)	6	.92
3.1 (Test r)	7	.88

2. Comparison of Bivariate Frequency Distributions of Test Score

The theoretical and observed bivariate distributions of scores on the pairs of tests investigated are compared in Tables 24 through 30. The same data with grouped cell frequencies are presented in Tables 31 through 37 in order to facilitate comparison of the theoretical and observed frequencies. The grouping was planned (without reference to the obtained frequencies) in such a way as to obtain theoretical frequencies of at least 9 cases in each cell, and also to maintain as regular a pattern of cells as possible without excessively coarse grouping. The dotted lines in the tables of ungrouped data indicate how the grouping was carried out.

As before, no significance test is possible with these data. Table 3, however, gives probabilities for the obtained values of chi-square that may be interpreted in the same fashion as the probabilities given in Table 2. It is seen that each of the theoretical distributions provides an adequate fit to the corresponding observed distribution.

Figures 12 through 18 present a graphic comparison of theoretical and observed regressions. The values for plotting these were obtained by computing the means of the columns and of the rows in Tables 24 through 30. The theoretical regression is shown only for the range of scores within which accurate regression values could be determined by means of the simple method of numerical integration used for this purpose. No further statistics relating to these regression lines have been calculated. Judging from the figures, however, we may say that the theoretical regressions seem to provide a good fit to the observed data.

Table 3

Probability that Certain Values of Chi-Square,
Having the Specified Degrees of Freedom, Will
Be Exceeded in Random Sampling.

<u>Chi-Square</u>	<u>Degrees of Freedom</u>	<u>Probability</u>
.6 (Tests 2 & 5)	6	.99
1.2 (Tests 2 & 8)	5	.94
7.2 (Tests 5 & 8)	6	.30
3.5 (Tests 5 & 82)	8	.90
5.4 (Tests 8' & r)	7	.62
4.8 (Tests 82 & r)	7	.68
3.3 (Tests h & L)	7	.86

IV. SUMMARY AND CONCLUSIONS

A number of important problems are generated by the fact that the peculiarities of the items in a test may sharply affect the shape of the frequency distribution and other properties of the test scores that we use as measures of the mental characteristics of a group of examinees. The present paper has developed a theory that, granted certain assumptions and restrictions, expresses the various properties and characteristics of test scores, in relation to each other and in relation to the underlying trait measured, as functions of the difficulties and intercorrelations of the items of which the test is composed.

The formulations and conclusions reached will be valid for any set of test data that is compatible with the assumptions and restrictions made. The theory developed facilitates and clarifies the interpretation to be placed on test scores as measures of a mental trait; it makes it possible to predict the properties of the test scores that will be obtained when a specified test is administered to a specified group of examinees; it indicates how items should be selected in order to construct tests that will have optimum properties for specific purposes.

Expressed in the language of achievement testing, the assumptions and restrictions underlying the development (outlined in Sections A1, 2, 3, and 4) were as follows:

- 1) All item responses are scored either 0 or 1.
- 2) The test score is the sum of the item scores.
- 3) The matrix of tetrachoric item intercorrelations has a rank of one when appropriate communalities are inserted in the diagonal.

- 4) The number of examinees is sufficiently large so that sampling fluctuations arising from sampling of examinees may be ignored.
- 5) There exists a measure of ability such that the probability of a correct answer to any item is a normal ogive function (see Figure 1) of the examinee's ability. This assumption, if it holds at all, provides the definition of the scale of measurement to be used for measuring ability. (This assumption implies that the test item cannot be answered correctly by guessing.)
- 6) The frequency distribution of ability in the group of examinees tested is Gaussian. (This assumption may be dispensed with in the majority of the derivations.)

The fifth and sixth assumptions may be replaced by the following single assumption (discussed in Section A5): The 2^n frequencies in the multivariate distribution of the item responses are such as could have arisen by the dichotomization of each of the variates in some normal multivariate population. A two-item test will always be compatible with these requirements. In the case of a longer test, it is always possible to test whether or not a given set of data is compatible with these requirements.

On this basis we have derived formulas for the bivariate distribution of test score and ability (Section B2), for the univariate distribution of test score (Section D1), for the univariate distribution of true score (Section E1), and for the bivariate distribution of scores on two tests of the same ability (Section F1). Various other expressions have been de-

rived, including those for the regression of test score on ability (Section B3), for the standard error of measurement at a given level of ability (Section B5), for the curvilinear correlation of test score on ability (Section B7), and for the frequency distribution of the errors of measurement for examinees having a given true score (Sections E1 and 2). In particular, an index was devised for measuring the discriminating power of the test at any specified level of ability (Section C).

Investigation of the formulas obtained lead to the following conclusions, among others:

- 1) The regression of test score on ability is necessarily curvilinear (Section B3, and also Figure 2). The actual effect of this curvilinearity may be negligible if the items are at a difficulty level appropriate to the group tested and if they are not too highly intercorrelated (Section B8); only in this case may the metric provided by the test scores be considered as providing units of measurement that correspond to equal units of ability, as here defined.
- 2) The standard error of measurement on a given test is ordinarily least for those examinees for whom the test is least discriminating, i.e., for those examinees who obtain near-zero or near-perfect scores (Section B5, and also the diagram for Test 2 in Figure 2).
- 3) The errors of measurement have binomial, not normal, distributions. Although they are uncorrelated with true score, they are not independent of true score, since the standard deviation

and skewness of their distribution varies depending on the true score (Sections E1 and E2).

- 4) Maximum discrimination at a given ability level, as defined by the discrimination index developed here, is provided by a test composed of items all of equal difficulty such that examinees at the given ability level will have a fifty per cent chance of answering each item correctly (Section C2).
- 5) There are strong indications, provided the item intercorrelations are not extraordinarily high, that a test composed entirely of items of fifty per cent difficulty will be more discriminating for practically all examinees in the group tested than will any test characterized by a spread of item difficulties. If the examinees have a normal distribution of ability, for example, the former test will be the most discriminating for all examinees except those who are more than, say, two-and-a-half standard deviations from the mean (Section C3, and also Figures 2 and 2A).
- 6) The shape of the frequency distribution of test scores or of true scores (Sections D and E, and also Figures 3-11) does not necessarily reflect the shape of the frequency distribution of ability. Sufficiently high tetrachoric item intercorrelations (.50 or higher) will produce rectangular or U-shaped distributions of test scores and of true scores even for groups having a normal distribution of ability. (The construction of tests that will produce rectangular or U-shaped score distribu-

tions has been urged more than once in the recent literature (7, 14); this goal can be approached, but its actual achievement, when the examinees have a normal distribution of ability, requires higher item intercorrelations than are at present usually obtained with most types of test items.)

The foregoing conclusions have here been derived only for the case where the test items cannot be answered correctly by guessing. The case where guessing occurs has been worked out in detail, but is not presented here.

An empirical study of seven short tests shows good agreement between the univariate and bivariate test score distributions actually obtained and the theoretical predictions of these distributions from the item difficulties and intercorrelations.

References

1. Brogden, H. Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. Psychometrika, 1946, 11, 197-214.
2. Carroll, J. B. Problems in the factor analysis of tests of varying difficulty. Amer. Psychologist, 1950, 5, 369. (Abstract.)
3. Cramér, H. Mathematical methods of statistics. Princeton: Princeton University Press, 1946.
4. Davis, F. B. Item-analysis data. Harvard University, Graduate School of Education, 1946.
5. Dunlap, J. W. and Kurtz, A. K. Handbook of statistical nomographs, tables, and formulas. Yonkers: World Book Company, 1932.
6. Ferguson, G. A. Item selection by the constant process. Psychometrika, 1942, 7, 19-29.
7. _____. On the theory of test discrimination. Psychometrika, 1949, 14, 61-68.
8. Finney, D. J. Probit analysis. Cambridge: Cambridge University Press, 1947.
9. Flanagan, J. C. General considerations in the selection of test items and a short method of estimating the product-moment coefficient from the data at the tails of the distributions. J. Educ. Psychol., 1939, 30, 674-680.

10. Guilford, J. P. Psychometric methods. New York and London: McGraw-Hill Book Company, 1936.
11. Gulliksen, H. The relation of item difficulty and inter-item correlation to test variance and reliability. Psychometrika, 1945, 10, 79-91.
12. _____. Theory of mental tests. New York: John Wiley and Sons; London: Chapman and Hall; 1950.
13. Hayes, S. P. Tables of the standard error of tetrachoric correlation coefficient. Psychometrika, 1943, 8, 193-203.
14. Jackson, R. W. B. and Ferguson, G. A. A functional approach in test construction. Educ. Psychol. Meas., 1943, 3, 23-28.
15. Kelley, T. L. Fundamentals of statistics. Cambridge: Harvard University Press, 1947, p. 370-373.
16. Kendall, M. G. The advanced theory of statistics. London: Charles Griffin and Company, 1945.
17. Kuder, G. F. Nomograph for point biserial r , biserial r , and four-fold correlations. Psychometrika, 1937, 2, 135-138.
18. Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.
19. Lawley, D. N. On problems connected with item selection and test construction. Proc. Roy. Soc. Edin., 1943, 61-A, Part 3, 273-287.
20. _____. The factorial analysis of multiple item tests. Proc. Roy. Soc. Edin., 1944, 62-A, Part I, 74-82.

21. Lazarsfeld, P. F. (with S. A. Stouffer et al). Measurement and prediction, Vol. 4 of studies in social psychology in World War II. Princeton: Princeton University Press, 1950, Chaps. 10 and 11.
22. Long, J. A. and Sandiford, P. The validation of test items. Bulletin No. 3, Department of Educational Research, University of Toronto, 1935.
23. Lorr, M. Interrelationships of number-correct and limen scores for an amount limit test. Psychometrika, 1944, 9, 17-30.
24. Mollenkopf, W. G. Variation of the standard error of measurement. Ph.D. thesis, Princeton University, 1948. Also Psychometrika, 1949, 14, 189-229.
25. Mosier, C. I. Psychophysics and mental test theory: fundamental postulates and elementary theorems. Psychol. Rev., 1940, 47, 355-366.
26. _____. Psychophysics and mental test theory. II. The constant process. Psychol. Rev., 1941, 48, 235-249.
27. Pearson, K. Mathematical contributions to the theory of evolution. Royal Soc. London, Phil. Trans., Series A, 1900, 195, 1-47.
28. _____. Tables for statisticians and biometricians. London: Cambridge University Press, 1924.
29. Peters, C. and Van Voorhis, W. Statistical procedures and their mathematical bases. New York and London: McGraw-Hill Book Company, 1940.

30. Plumlee, L. B. An investigation of the effect of difficulty and chance success on item-test correlation and on test reliability. Ph.D. thesis, University of Chicago, 1950.
31. Richardson, M. W. Relation between the difficulty and the differential validity of a test. Ph.D. thesis, University of Chicago, 1936. Also Psychometrika, 1936, 1 (No. 2), 33-49.
32. Symonds, P. M. Choice of items for a test on the basis of difficulty. J. Educ. Psychol., 1928, 19, 73-87.
33. Thorndike, R. L. Personnel selection. New York: John Wiley and Sons, 1949, pp. 228-230.
34. Thurstone, T. The difficulty of a test and its diagnostic value. J. Educ. Psychol., 1932, 23, 335-43.
35. Tucker, L. R. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-13.
36. _____. A method for scaling ability test items in difficulty taking item unreliability into account. Amer. Psychologist, 1948, 3, 309-10. (Abstract.)
37. Wherry, R. J. and Gaylord, R. H. Factor pattern of test items and tests as a function of the correlation: content, difficulty, and constant error factors. Psychometrika, 1944, 9, 237-44.
38. Wilks, S. S. Mathematical statistics. Princeton: Princeton University Press, 1944.

APPENDICES

	<u>Page</u>
Figures 4-18	81
Tables 4-37	93
Summary of Notation (Appendix A)	1A

F I G U R E S

	<u>Figs.</u>
Theoretical and Observed Distributions of Scores	4-11
Theoretical and Actual Regressions	12-18

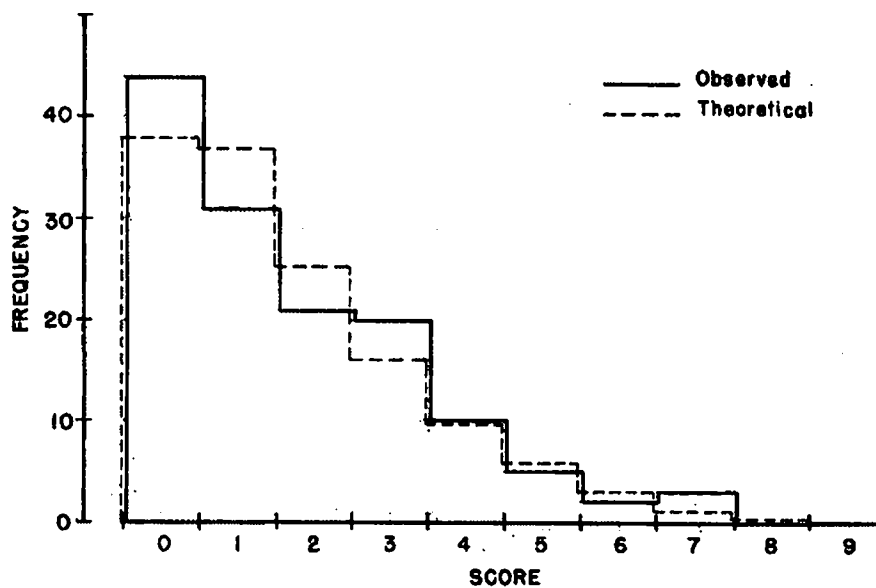


FIGURE 4

Theoretical and Observed Distributions
of Scores on Test 2 (difficult items).

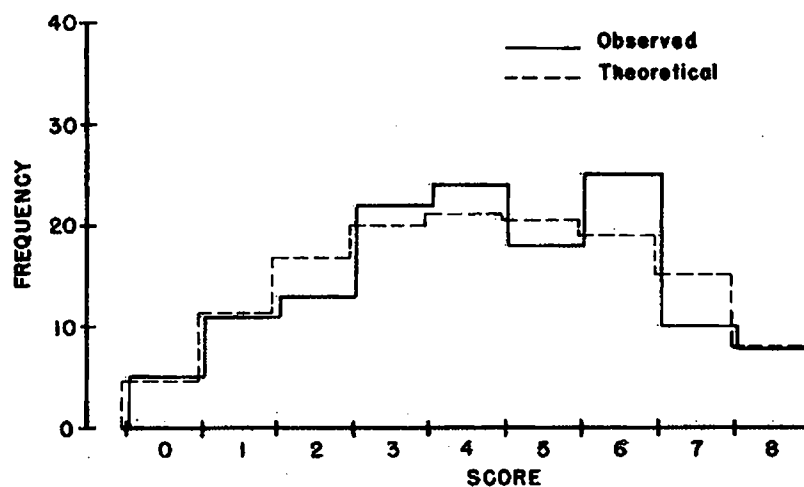


FIGURE 5

Theoretical and Observed Distributions
of Scores on Test 5 (items of medium difficulty).

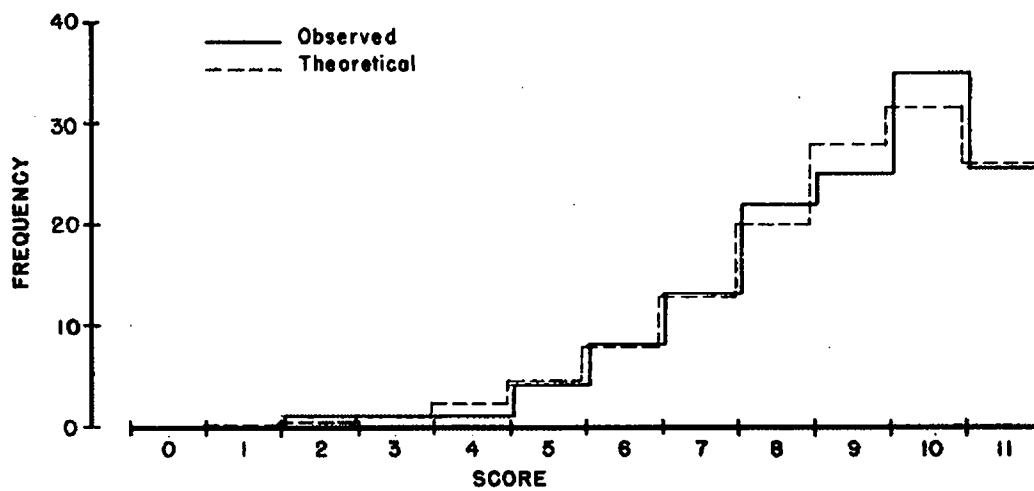


FIGURE 6

Theoretical and Observed Distributions
of Scores on Test 8 (easy items).

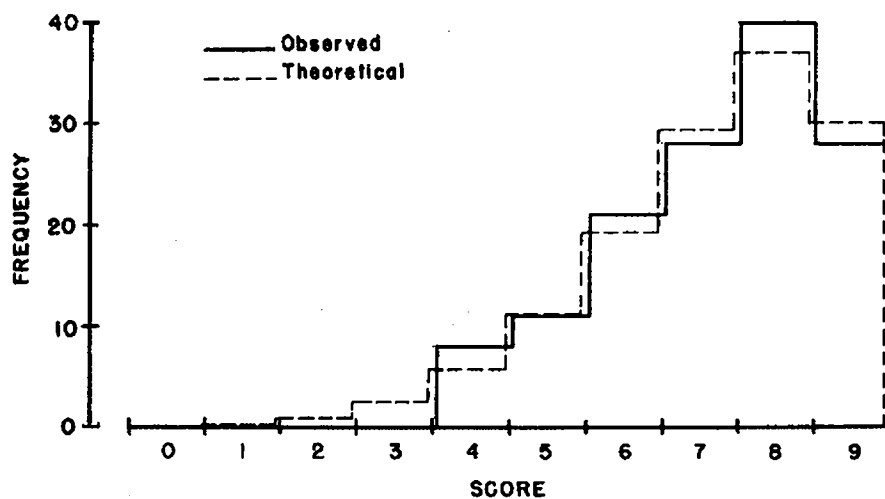


FIGURE 7

Theoretical and Observed Distributions
of Scores on Test 8' (easy items).

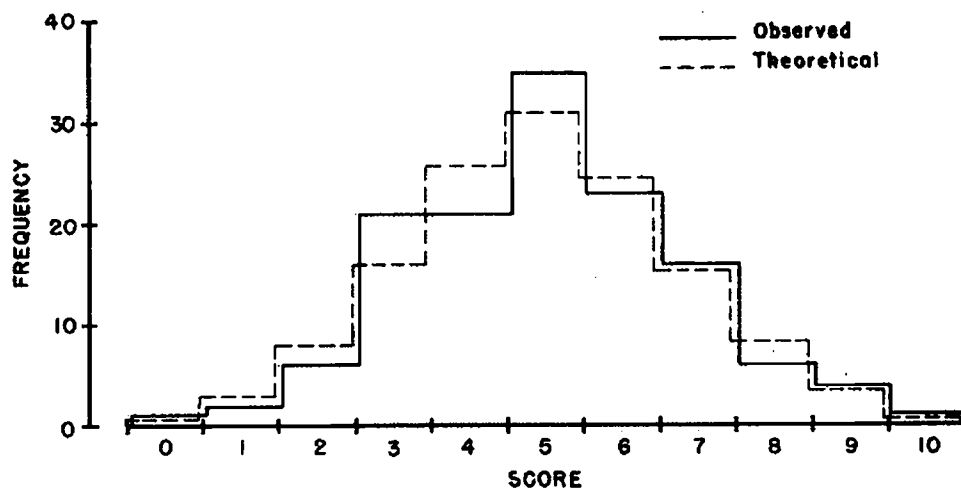


FIGURE 8

Theoretical and Observed Distributions of Scores on Test 82 (five easy and five difficult items).

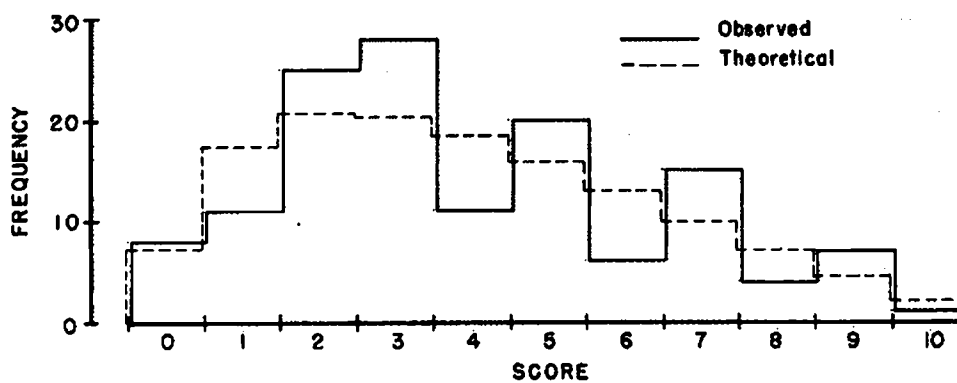


FIGURE 9

Theoretical and Observed Distributions of Scores on Test h (highly discriminating items).

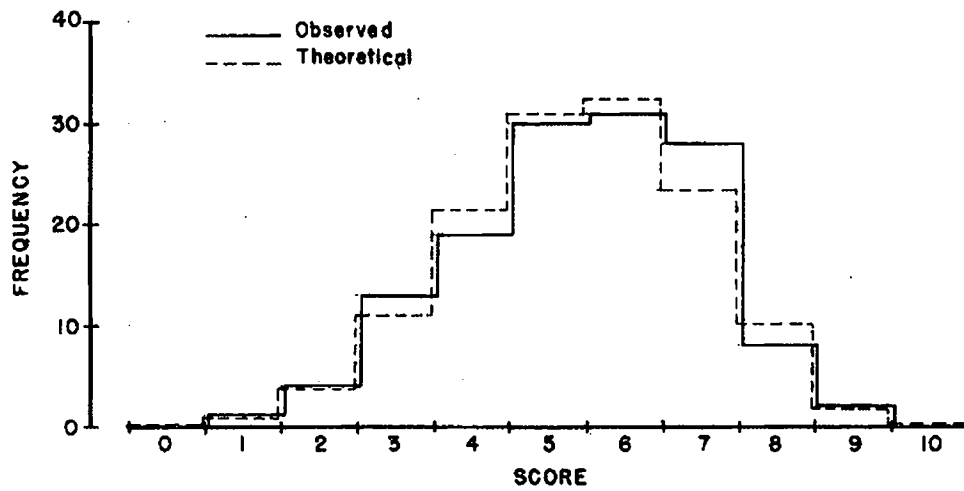


FIGURE 10

Theoretical and Observed Distributions of Scores
on Test L (poorly discriminating items).

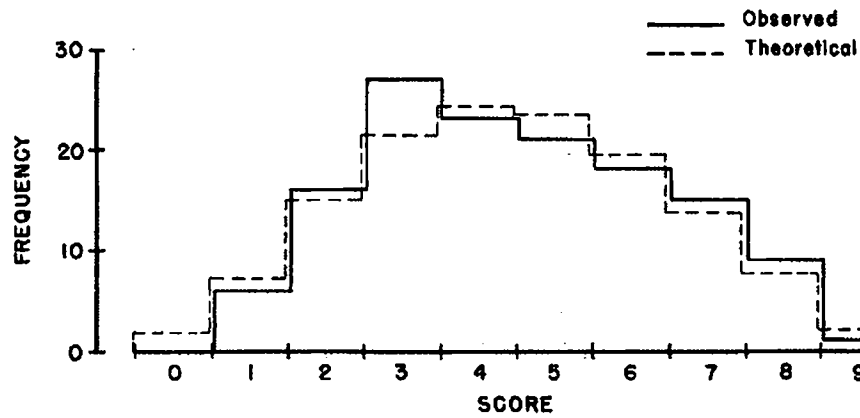


FIGURE 11

Theoretical and Observed Distributions of Scores
on Test r (rectangular distribution of item difficulty).

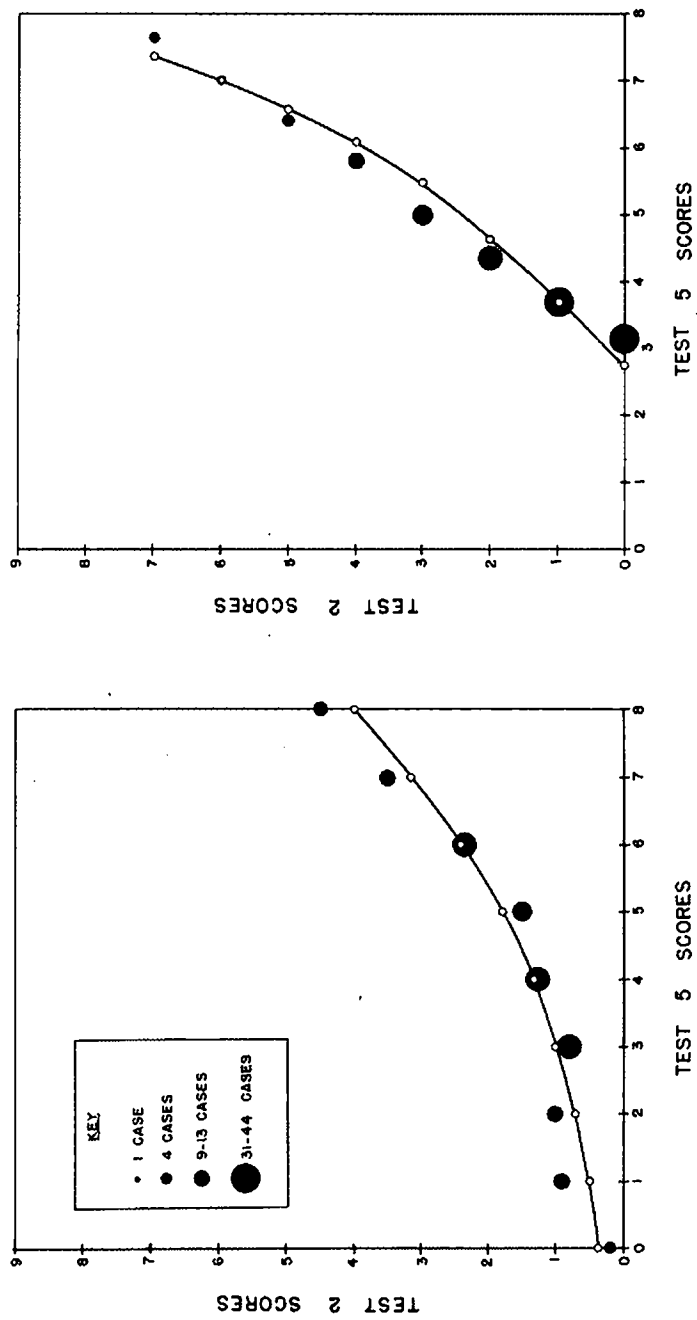


FIGURE 12
Theoretical (hollow circles) and Actual
(solid circles) Regressions for Test 2 on
Test 5 (left) and for Test 5 on Test 2 (right)

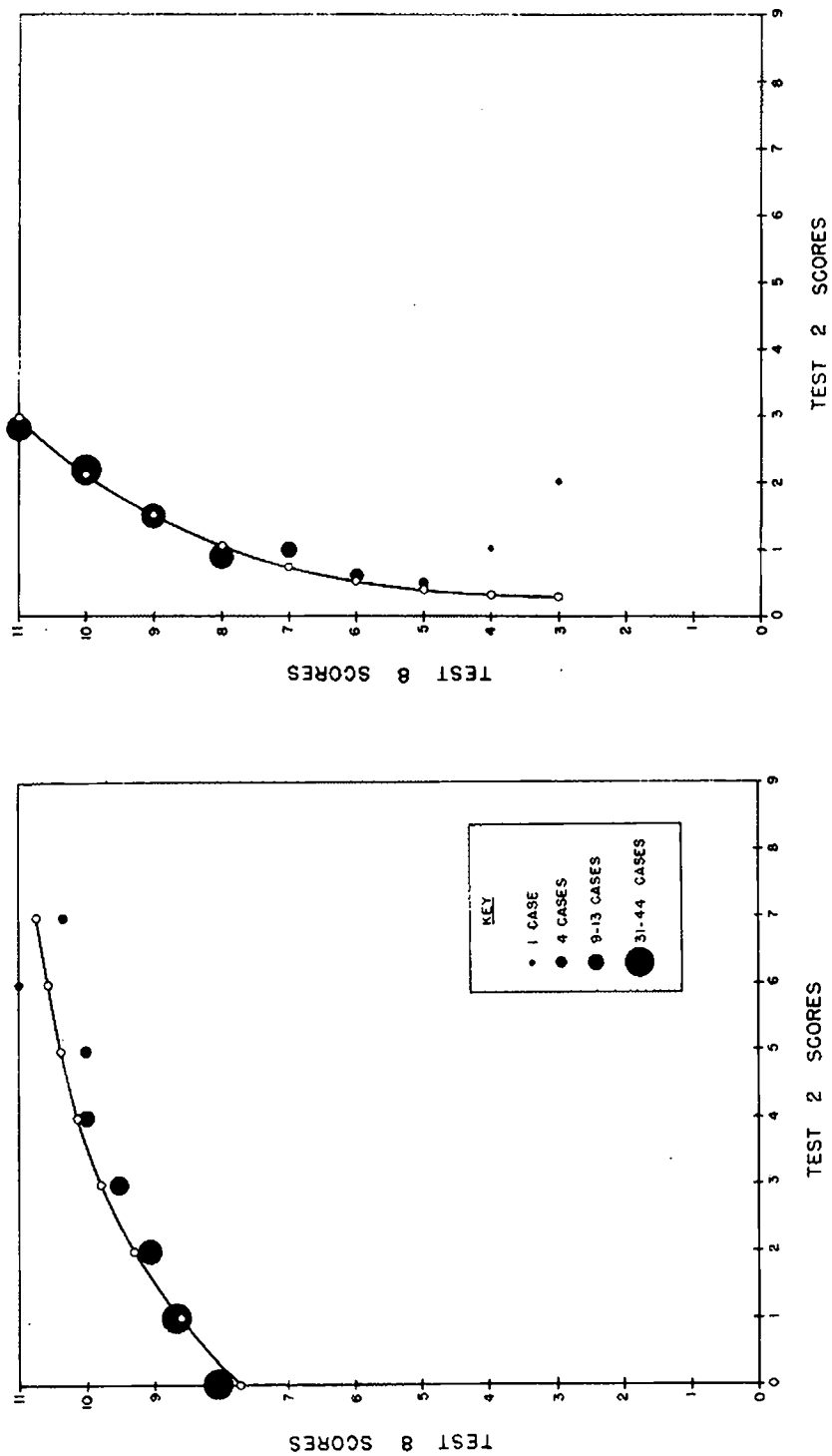


FIGURE 13
Theoretical (hollow circles) and Actual (solid circles) Regressions for Test 8 on Test 2 (left) and for Test 2 on Test 8 (right).

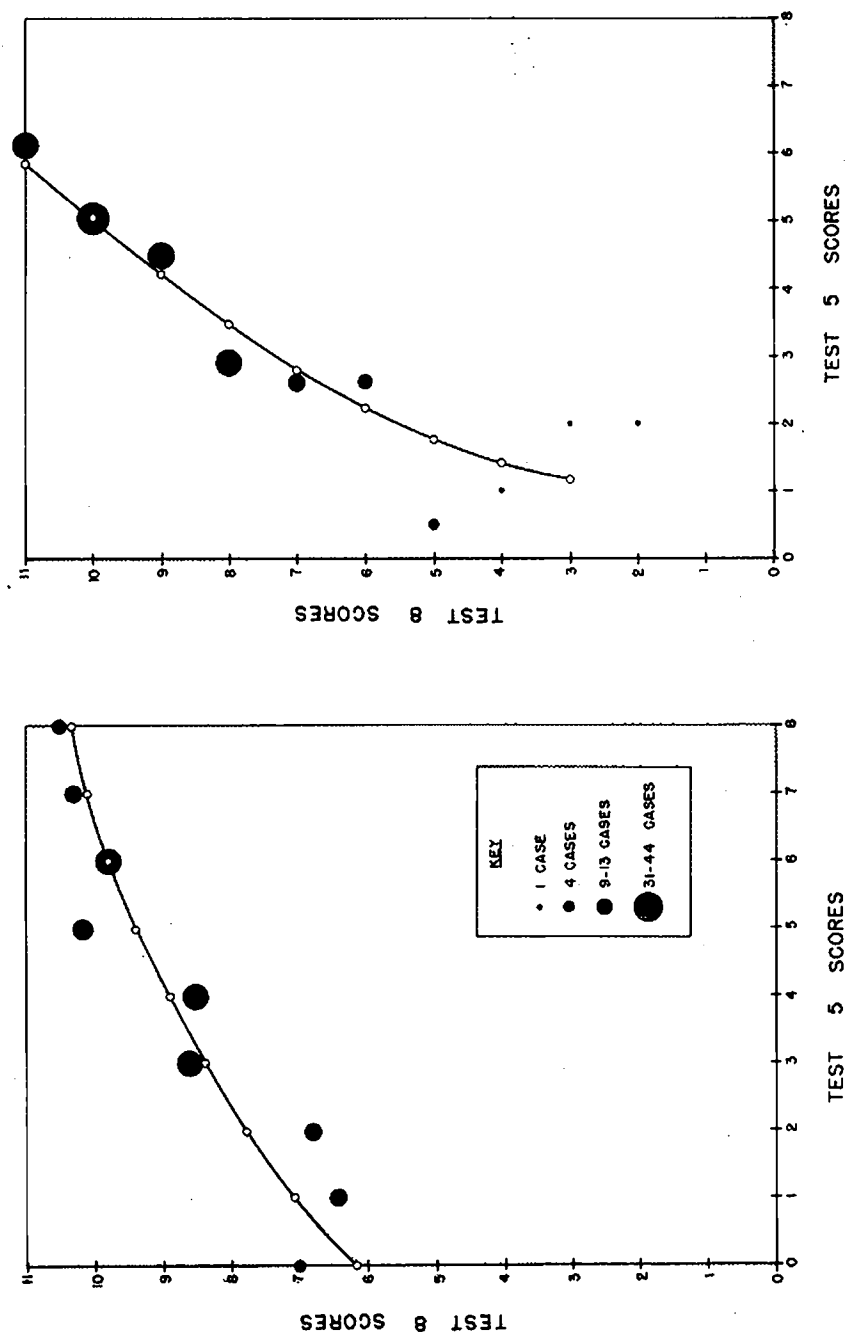


FIGURE 14
Theoretical (hollow circles) and Actual (solid circles) Regressions for Test 8 on Test 5 (left) and for Test 5 on Test 8 (right).

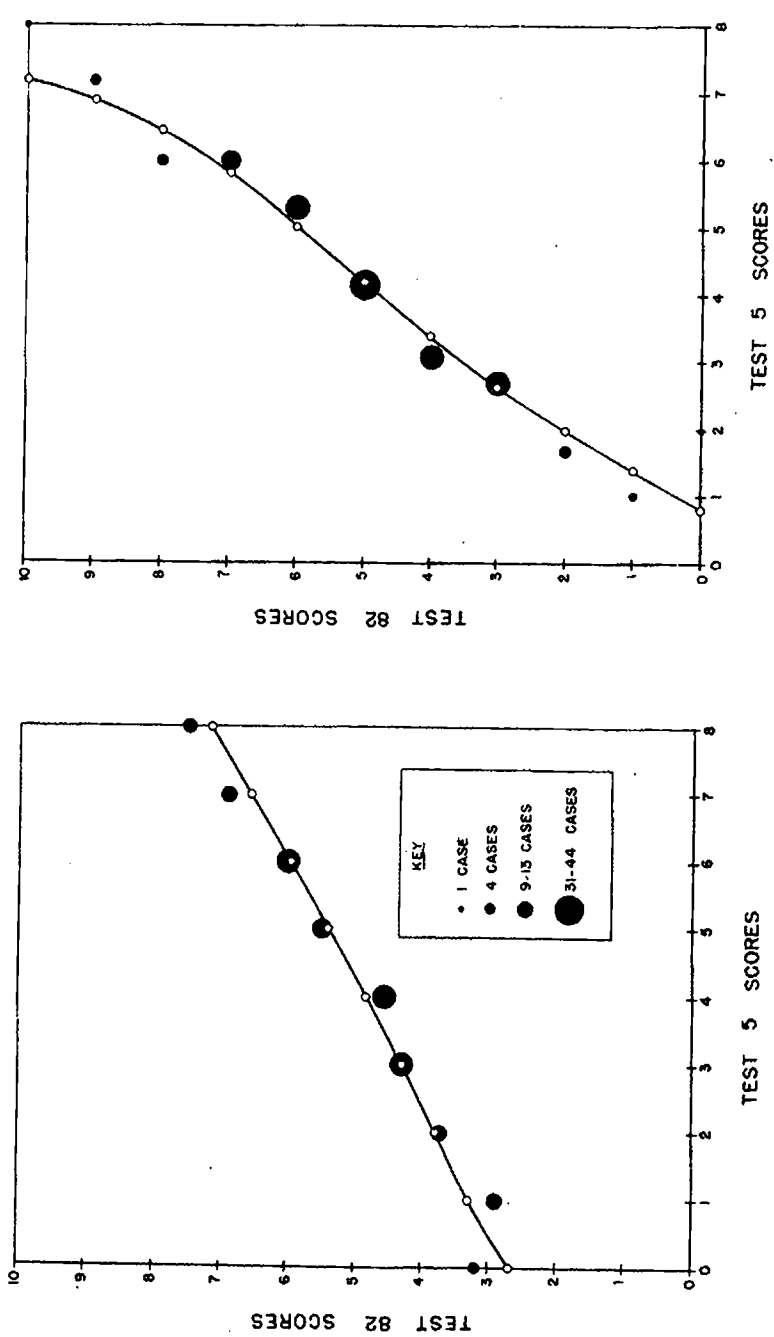


FIGURE 15
Theoretical (hollow circles) and Actual
solid circles) Regressions for Test 82 on
Test 5 (left) and for Test 5 on Test 82 (right).

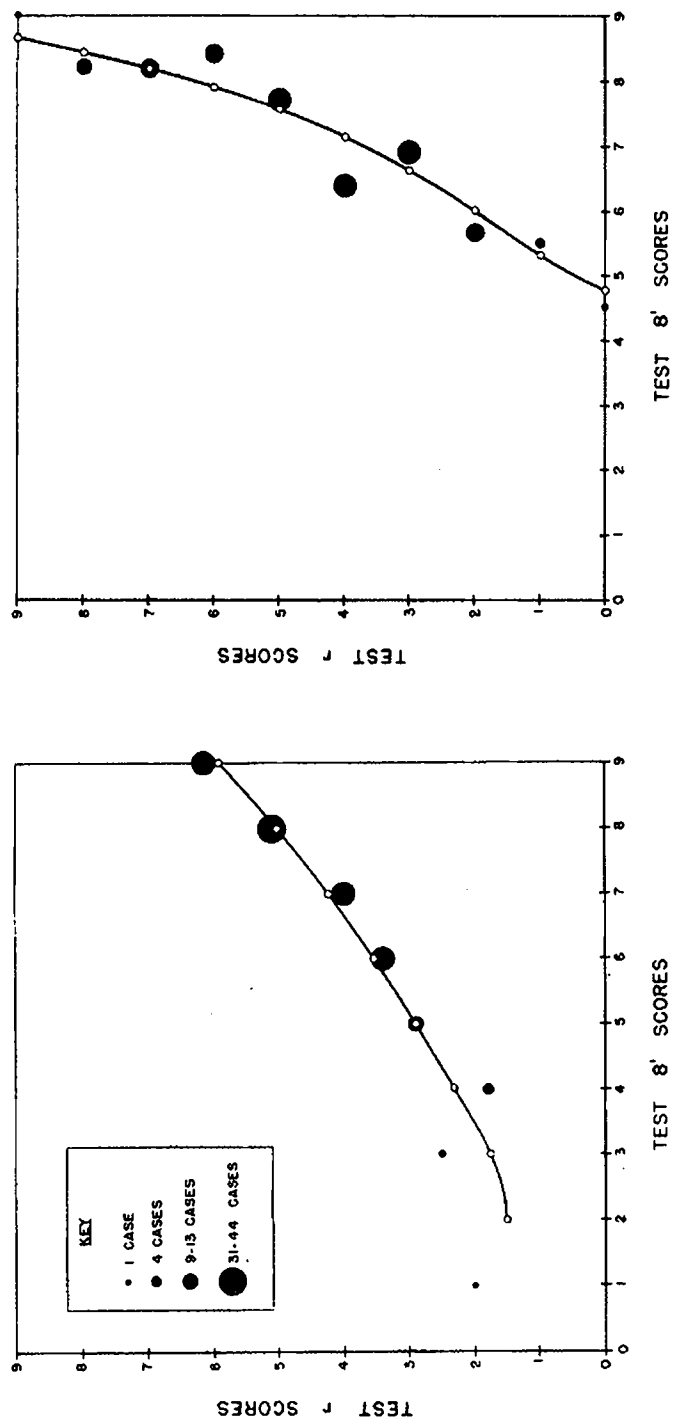


FIGURE 16
Theoretical (hollow circles) and Actual
(solid circles) Regressions for Test r on
Test 8, left and for Test 8, on Test r (right).

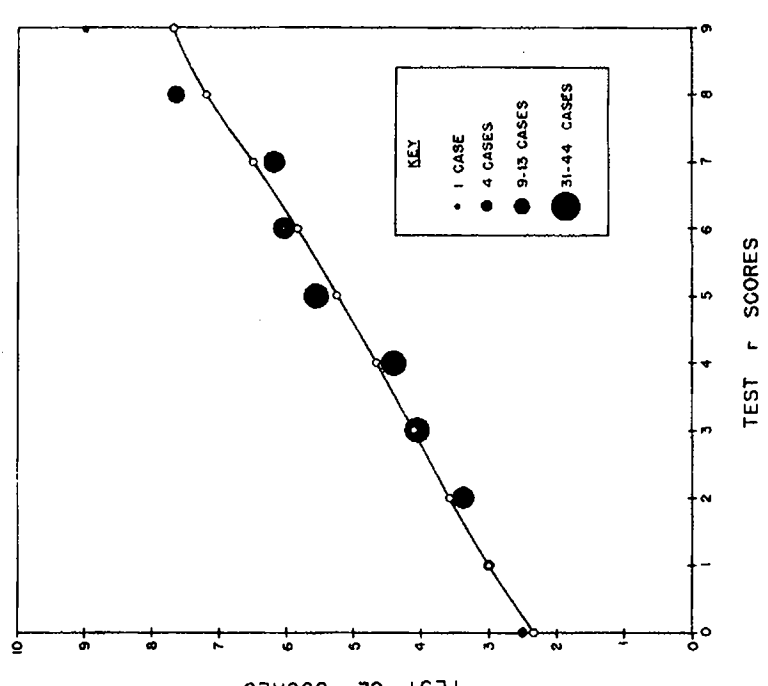
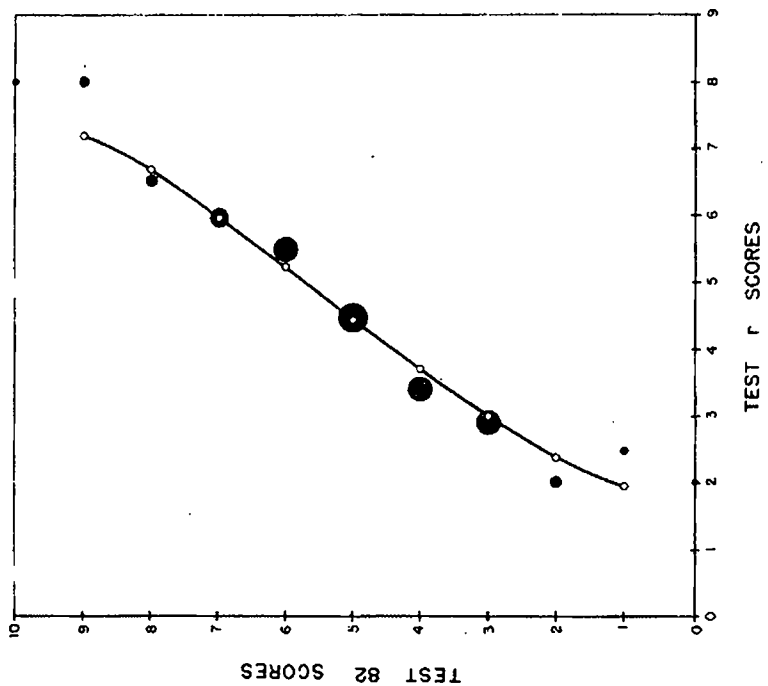


FIGURE 17

Theoretical (hollow circles) and Actual (solid circles) Regressions for Test 82 on Test r (left) and for Test r on Test 82 (right).

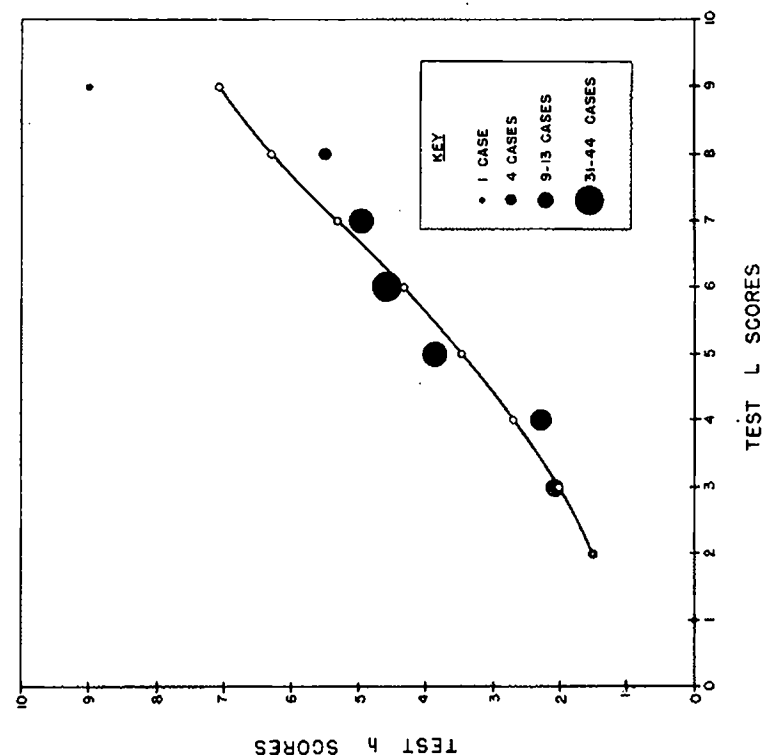
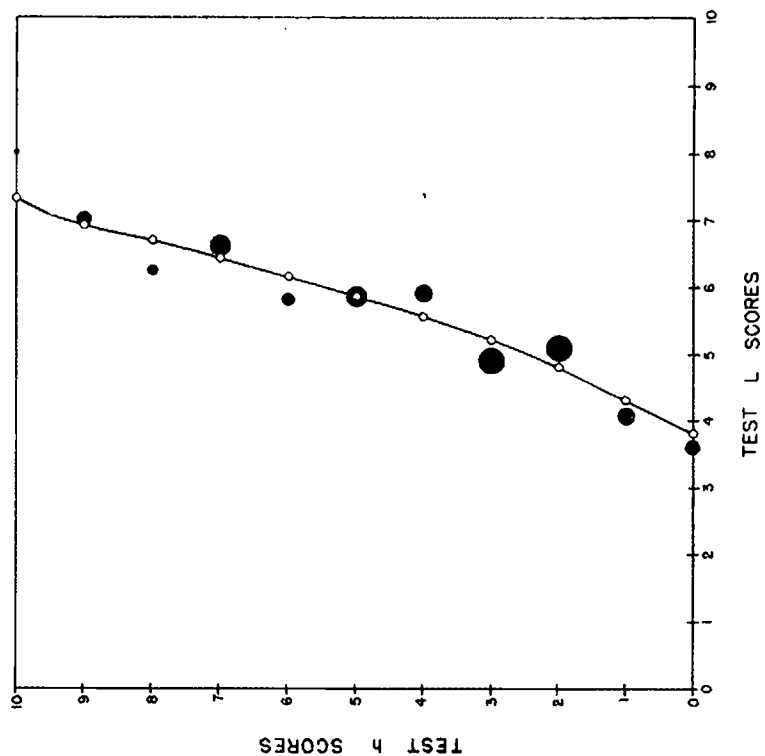


FIGURE 18
 Theoretical (hollow circles) and Actual
 (solid circles) Regressions for Test h on
 Test L (left) and for Test L on Test h (right).

T A B L E S

	<u>Tables</u>
Tetrachoric Intercorrelations among Twenty-Eight Items	4
The Basic Item Statistics: Item Difficulties and Common Factor Loadings	5
Residual Correlations among Twenty-Eight Items after Extraction of the First Factor	6
Item Difficulties and Common Factor Loadings for Each of the Eight Tests	7
Bivariate Frequency Distribution of Ability and Score; Also Marginal Distribution of Test Score	8-15
Comparison of Theoretical and Actual Score Distributions . . .	16-23
Scatterdiagrams for Pairs of Tests, Showing Actual Frequencies and Theoretical Frequencies	24-30
Bivariate Grouped Frequency Tables for Pairs of Tests, Showing Observed Frequencies and Theoretical Frequencies	31-37

Table 4

Tetrachoric Intercorrelations Among Twenty-Eight Items (decimal points omitted).

Item No.	1'	3'	5'	1	2	3	5	6	8	9	10	11	13	16	18	20	39	40	44	45	46	48	50	53	54	55	56	57
1'	33	33	50	26	26	29	40	*	36	50	62	20	29	10	19	22	02	56	26	41	24	36	27	69	-08	*	08	37
3'	50	28	28	15	15	52	46	28	16	51	34	52	12	22	28	-10	09	05	18	43	14	54	24	38	34	28	16	10
5'	44	15	44	15	24	24	53	30	18	38	38	25	56	22	22	40	-05	32	51	42	10	34	04	36	-06	12	33	26
1	26	15	15	32	27	27	*	*	-02	03	30	47	25	41	14	19	-16	51	43	31	27	12	32	41	-30	31	25	04
2	26	15	15	32	20	20	*	*	22	-08	20	30	13	41	14	-02	24	30	24	21	27	34	51	41	-13	*	43	35
3	29	52	24	27	20	37	37	15	36	52	22	25	-03	00	*	19	-18	21	25	24	17	41	29	48	-02	48	10	10
5	40	46	53	*	*	*	37	20	36	27	59	53	21	33	33	36	10	28	50	06	19	45	36	51	-05	42	56	20
6	*	28	30	*	*	*	20	10	10	30	42	58	07	25	-08	22	10	25	24	38	29	48	10	36	*	14	12	41
8	36	16	18	-02	22	36	36	10	10	10	15	19	17	-06	-07	07	04	11	17	01	28	08	15	26	05	30	17	19
9	50	51	38	03	-08	52	27	30	10	15	34	34	38	30	30	-01	23	00	24	32	16	47	-05	32	08	33	13	30
10	62	34	38	30	20	22	59	42	10	15	34	34	17	09	28	-09	10	52	28	24	39	42	19	44	-11	45	35	21
11	20	52	25	47	30	25	53	21	19	19	38	17	17	23	08	08	18	37	33	12	18	48	33	52	04	14	38	37
13	29	12	22	25	13	-03	21	07	17	17	38	17	23	02	02	09	03	15	35	22	30	10	27	13	-02	20	23	09
16	10	22	22	41	41	*	33	25	10	00	30	34	01	52	10	07	10	25	10	39	18	26	30	13	-06	41	12	23
18	19	28	22	17	08	08	09	02	07	07	16	16	08	09	07	16	16	21	24	07	08	22	43	39	01	17	28	02
20	02	09	05	14	-02	19	36	22	07	-01	34	-09	08	09	10	16	11	11	00	-08	-01	30	17	23	08	-12	-29	14
39	02	09	05	14	-02	19	36	22	07	-01	34	-09	08	09	10	16	11	11	00	-08	-01	30	17	23	08	-12	-29	14
40	56	05	32	21	24	25	28	24	11	00	52	28	37	35	25	21	00	33	33	40	20	40	32	45	21	33	39	19
44	26	18	51	43	24	24	28	24	17	24	28	24	33	35	10	24	00	33	05	-05	26	56	33	35	-28	26	33	17
45	41	43	42	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26
46	41	43	42	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26
48	41	43	42	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26
50	41	43	42	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26
53	41	43	42	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26
54	41	43	42	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26
55	41	43	42	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26
56	41	43	42	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26
57	41	43	42	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26

*Correlations are not recorded in cases where there is zero frequency in one cell of the fourfold table.

Table 5

The Basic Item Statistics: Item Difficulties (p_i)
and Common Factor Loadings (R_i).

Item No.	p_i	R_i
1 ^c	.868	.654
3 ⁱ	.338	.549
5 ⁱ	.471	.595
1	.882	.481
2	.882	.495
3	.676	.476
5	.199	.717
6	.096	.490
8	.801	.299
9	.272	.409
10	.574	.640
11	.272	.581
13	.809	.466
16	.522	.352
18	.096	.332
20	.735	.324
39	.897	.114
40	.434	.593
44	.713	.531
45	.750	.485
46	.485	.437
48	.176	.715
50	.801	.530
53	.441	.800
54	.088	-.093
55	.162	.655
56	.610	.483
57	.721	.465

Table 6

Residual Correlations Among 28 Items After Extraction of the First Factor
(decimal points omitted).

Item No.	1'	3'	5'	1	2	3	5	6	8	9	10	11	13	16	18	20	39	40	44	45	46	48	50	53	54	55	56	57
1'																												
3'	-03																											
5'	11	-05																										
1	-06	-11	15																									
2	-06	-12	-14	08																								
3	-02	26	-04	04	-04																							
5	-07	07	10	*	*	03																						
6	*	01	01	*	*	-08	-15																					
8	16	-00	00	-16	07	22	15	-05																				
9	23	28	14	-17	-28	32	-02	10	-02																			
10	20	-01	-00	-01	-12	-08	13	11	-15	-11																		
11	-18	20	-10	19	01	-03	11	30	02	-05	-03																	
13	-02	-14	28	03	-10	-25	-12	-16	03	-02	08	-10																
16	-13	03	01	24	24	-17	08	08	-16	-14	08	-12	07															
18	-03	10	02	-02	-02	*	09	-24	-17	-21	09	-08	-10	-10	-12	-04												
20	01	-28	21	03	-18	04	13	06	-03	-14	13	-28	-07	-02	-04													
39	-06	03	-12	-22	18	-23	02	04	01	18	-06	03	13	-01	06	12												
40	17	-28	-03	22	01	-07	-14	-04	-07	-24	14	-06	09	-06	05	02	04											
44	-09	-11	19	18	-02	-00	12	-02	01	02	-06	-07	08	16	-08	07	-06	02										
45	09	16	13	08	-03	01	-29	14	-14	12	-04	-04	-11	05	23	-09	-14	11	-31									
46	-05	-10	-16	06	05	-04	-12	08	15	-02	-04	14	-02	15	04	-06	-06	-06	09	05								
48	-11	15	-08	-22	-01	07	-06	13	-13	07	01	00	15	-15	02	-01	22	-02	-12	21	-28							
50	-08	-05	-28	06	25	04	-02	-16	-01	-27	-11	-12	08	08	12	26	11	01	10	07	-11	11						
53	17	-06	-12	02	01	10	-06	-03	02	-01	-07	04	15	-15	-14	13	14	-02	-04	-04	05	01	-03					
54	-02	39	-00	-26	-08	02	02	*	08	12	-05	-12	08	01	-03	04	09	26	-26	-24	-07	22	-18	03				
55	*	-08	-27	-00	*	17	-05	-18	10	06	03	-05	-16	-03	19	-04	-20	-06	08	-06	18	10	11	01	06			
56	-24	-10	40	02	19	-13	21	-12	03	-07	40	10	10	06	-04	12	-34	10	-05	10	-13	80	28	13	10	60		
57	07	-16	-20	-18	12	-12	-13	18	05	11	-09	10	-10	-07	08	-13	09	-09	16	-06	30	-80	-28	10	-03	12	60	

*Residuals are not recorded in cases where there is zero frequency in one cell of the fourfold table. The nine unrecorded residuals are .68, .57, .66, .76, .64, .76, .68, .84, and .95.

Table 7

**Item Difficulties and Common Factor
Loadings for Each of the Eight Tests.**

Item No.	Item Diffi- culty	Common Factor Loadings of Test Items							
		Test 2	Test 5	Test 8	Test 8'	Test 82	Test h	Test L	Test r
54	.088	-.09	--	--	--	--	--	-.09	--
6	.096	.49	--	--	--	--	--	--	.49
18	.096	.33	--	--	--	.33	--	.33	--
55	.162	.66	--	--	--	.66	.66	--	--
48	.176	.72	--	--	--	.72	.72	--	--
5	.199	.72	--	--	--	--	.72	--	.72
9	.272	.41	--	--	--	.41	--	.41	--
11	.272	.58	--	--	--	.58	.58	--	--
3'	.338	.55	--	--	--	--	.55	--	.55
40	.434	--	.59	--	--	--	.59	--	.59
53	.441	--	.80	--	--	--	.80	--	--
5'	.471	--	.60	--	--	--	.60	--	.60
46	.485	--	.44	--	--	--	--	.44	--
16	.522	--	.35	--	--	--	--	.35	--
10	.574	--	.64	--	--	--	.64	--	.64
56	.610	--	.48	--	--	--	--	--	--
3	.676	--	.48	--	--	--	--	--	.48
44	.713	--	--	.53	.53	.53	--	--	--
57	.721	--	--	.46	.46	--	--	.46	--
20	.735	--	--	.32	.32	--	--	.32	--
45	.750	--	--	.48	.48	.48	--	--	--
50	.801	--	--	.53	--	--	--	--	.53
8	.801	--	--	.30	.30	--	--	.30	--
13	.809	--	--	.47	.47	.47	--	.47	--
1'	.868	--	--	.65	.65	.65	.65	--	--
1	.882	--	--	.48	.48	.48	--	--	--
2	.882	--	--	.50	--	--	--	--	.50
39	.897	--	--	.11	.11	--	--	.11	--
No. of items:		9	8	11	9	10	10	10	9

Table 8

Bivariate Frequency Distribution of Ability and Score
on Test 2 (Difficult Items) Showing Frequency per 10,000;
Also the Marginal Distribution of Test Score (f_g).

Test Score	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	f_s
9	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
8	--	--	--	--	--	--	--	--	--	1	4	9	10	6	2	.002
7	--	--	--	--	--	--	--	--	1	11	36	54	40	16	4	.008
6	--	--	--	--	--	--	--	1	14	66	140	132	61	15	2	.022
5	--	--	--	--	--	--	1	14	83	237	298	166	44	6	1	.043
4	--	--	--	--	--	2	15	94	319	519	373	117	17	1	--	.103
3	--	--	--	--	4	24	120	405	768	703	283	48	4	--	--	.118
2	--	--	2	11	51	194	557	1057	1122	575	128	12	--	--	--	.181
1	1	7	32	115	343	814	1392	1514	906	259	31	2	--	--	--	.122
0	7	37	142	414	897	1386	1435	905	307	49	3	--	--	--	--	.279

Table 9

Bivariate Frequency Distribution of Ability and Score
on Test 5 (Items of Medium Difficulty) Showing Frequency per
10,000; Also the Marginal Distribution of Test Score (f_s).

Test Score	A b i l i t y S c o r e (<u>a</u>)															f _s
	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	
8	--	--	--	--	--	--	1	20	118	289	351	244	108	33	7	.059
7	--	--	--	--	--	1	20	158	508	720	512	210	55	10	1	.110
6	--	--	--	--	1	13	124	524	945	763	309	72	11	1	--	.138
5	--	--	--	--	8	81	414	972	988	448	101	13	1	--	--	.151
4	--	--	--	5	51	288	830	1103	634	159	19	1	--	--	--	.155
3	--	--	4	34	191	614	1020	783	255	35	2	--	--	--	--	.147
2	--	3	23	121	401	765	753	339	63	5	--	--	--	--	--	.124
1	2	14	68	220	444	515	306	82	9	--	--	--	--	--	--	.083
0	6	27	80	159	199	143	52	8	1	--	--	--	--	--	--	.034

Table 10

Bivariate Frequency Distribution of Ability and Score
on Test 8 (Easy Items) Showing Frequency per 10,000;
Also the Marginal Distribution of Test Score (f_s).

Test Score	A b i l i t y S c o r e (<u>c</u>)															f _s
	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	
11	--	--	--	--	3	31	173	515	903	992	725	370	136	37	8	.195
10	--	--	--	2	25	174	614	1182	1329	934	438	144	35	6	1	.244
9	--	--	1	11	96	427	974	1205	865	386	114	24	4	--	--	.205
8	--	--	3	37	213	615	905	719	328	92	17	2	--	--	--	.147
7	--	--	11	82	309	576	548	278	80	14	2	--	--	--	--	.095
6	--	1	25	125	305	369	227	73	13	1	--	--	--	--	--	.057
5	--	3	41	130	208	165	65	14	1	--	--	--	--	--	--	.032
4	1	7	45	93	98	51	13	2	--	--	--	--	--	--	--	.016
3	2	11	32	45	31	11	2	--	--	--	--	--	--	--	--	.007
2	3	12	14	13	6	1	--	--	--	--	--	--	--	--	--	.002
1	2	8	3	2	1	--	--	--	--	--	--	--	--	--	--	.001
0	1	3	3	2	1	--	--	--	--	--	--	--	--	--	--	---

Table 11
 Bivariate Frequency Distribution of Ability and Score
 on Test 8' (Easy Items) Showing Frequency per 10,000;
 Also the Marginal Distribution of Test Score (f_g).

Test Score	A b i l i t y S c o r e (<u>c</u>)															f _s
	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	
9	--	--	--	1	8	62	266	671	1050	1075	754	376	137	37	8	.222
8	--	--	--	6	58	292	818	1349	1381	924	425	140	34	6	1	.272
7	--	--	3	30	181	592	1088	1174	783	340	101	22	3	--	--	.216
6	--	1	12	82	319	681	820	578	251	70	13	2	--	--	--	.141
5	--	4	30	136	347	488	386	177	50	9	1	--	--	--	--	.081
4	1	9	48	144	242	225	117	35	6	1	--	--	--	--	--	.041
3	2	14	47	95	107	67	23	4	--	--	--	--	--	--	--	.018
2	3	12	27	37	29	12	3	--	--	--	--	--	--	--	--	.006
1	2	5	8	8	4	1	--	--	--	--	--	--	--	--	--	.001
0	--	1	1	1	--	--	--	--	--	--	--	--	--	--	--	---

Table 12
Bivariate Frequency Distribution of Ability and Score on
Test 82 (Both Easy and Difficult Items) Showing Frequency per
10,000; Also the Marginal Distribution of Test Score (f_s).

Test Score	A b i l i t y S c o r e ($\bar{c}$)															f_s
	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	
10	--	--	--	--	--	--	--	--	1	7	20	28	21	10	3	.005
9	--	--	--	--	--	--	--	3	22	82	149	137	71	22	5	.025
8	--	--	--	--	--	--	4	38	167	354	371	200	59	11	1	.060
7	--	--	--	--	--	7	58	264	620	730	429	128	20	2	--	.113
6	--	--	--	1	10	82	369	906	1170	766	249	40	3	--	--	.180
5	--	--	1	10	88	409	1052	1461	1063	390	70	6	--	--	--	.228
4	--	--	6	56	283	787	1193	964	401	82	8	--	--	--	--	.189
3	--	3	27	143	432	718	644	302	71	8	--	--	--	--	--	.117
2	1	11	58	185	335	334	176	47	6	--	--	--	--	--	--	.058
1	4	19	60	117	128	76	23	4	--	--	--	--	--	--	--	.022
0	4	12	24	28	19	7	1	--	--	--	--	--	--	--	--	.005

Table 13
Bivariate Frequency Distribution of Ability and Score
on Test h (Highly Valid Items) Showing Frequency per
10,000; Also the Marginal Distribution of Test Score (f_s).

Test Score	A b i l i t y S c o r e (<u>c</u>)															f _s
	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	
10	--	--	--	--	--	--	--	--	1	14	63	104	77	30	7	.015
9	--	--	--	--	--	--	--	--	14	107	243	197	68	12	1	.032
8	--	--	--	--	--	--	--	7	96	355	395	155	25	2	--	.052
7	--	--	--	--	--	--	2	55	353	634	347	66	5	--	--	.073
6	--	--	--	--	--	1	24	250	763	670	180	16	1	--	--	.095
5	--	--	--	--	--	10	143	683	1001	430	56	2	--	--	--	.116
4	--	--	--	--	6	82	503	1130	801	167	10	--	--	--	--	.135
3	--	--	--	4	56	361	1033	1108	380	38	1	--	--	--	--	.149
2	--	--	5	48	282	847	1155	600	99	5	--	--	--	--	--	.152
1	--	7	51	233	622	889	588	148	12	--	--	--	--	--	--	.128
0	8	37	119	254	330	230	72	8	--	--	--	--	--	--	--	.053

Table 14

Bivariate Frequency Distribution of Ability and Score
on Test L (Low Validity Items) Showing Frequency per
10,000; Also the Marginal Distribution of Test Score (f_s).

Test Score	A b i l i t y S c o r e (<u>c</u>)															f _s
-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5		
10	--	--	--	--	--	--	1	2	3	3	2	1	--	--		.001
9	--	--	--	--	1	5	18	39	60	63	47	25	9	3		.014
8	--	--	--	3	15	61	165	297	363	303	175	71	20	4		.074
7	--	--	--	23	104	316	638	858	772	470	195	56	11	2		.172
6	--	--	3	21	356	795	1172	1140	737	319	94	19	3	--		.238
5	--	2	13	73	269	662	1073	795	366	112	23	3	--	--		.227
4	1	6	36	145	390	696	815	310	101	22	3	--	--	--		.157
3	2	13	56	164	323	419	356	70	16	2	--	--	--	--		.081
2	3	15	47	101	147	141	87	9	1	--	--	--	--	--		.029
1	2	7	18	30	33	24	11	3	--	--	--	--	--	--		.006
0	--	1	2	3	3	1	1	--	--	--	--	--	--	--		.001

Table 15

Bivariate Frequency Distribution of Ability and Score on Test r (Rectangular Distribution) Showing Frequency per 10,000; Also the Marginal Distribution of Test Score (f_s).

Test Score	Ability Score (c)														f _s	
	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	
9	--	--	--	--	--	--	--	--	7	36	81	90	56	22	6	.015
8	--	--	--	--	--	--	1	16	104	288	368	238	87	20	3	.056
7	--	--	--	--	--	1	20	149	487	707	475	158	28	3	--	.101
6	--	--	--	--	1	18	152	584	1009	780	275	46	4	--	--	.143
5	--	--	--	1	13	124	551	1141	1070	445	82	7	--	--	--	.172
4	--	--	1	10	91	438	1052	1197	616	138	13	1	--	--	--	.178
3	--	1	8	65	309	802	1065	677	193	23	1	--	--	--	--	.157
2	1	6	41	182	485	716	545	197	31	2	--	--	--	--	--	.110
1	3	19	79	207	324	282	126	26	2	--	--	--	--	--	--	.053
0	5	19	47	76	72	38	10	1	--	--	--	--	--	--	--	.013

Score	Theoretical Frequency (F_t)	Actual Frequency (F_o)	$\frac{(F_o - F_t)^2}{F_t}$
9	--	--	.00
8	0.3	--	
7	1.1	3	
6	3.0	2	
5	5.8	5	
4	9.9	10	.00
3	16.0	20	1.00
2	25.2	21	.70
1	36.9	31	.94
0	37.9	44	.98
M	1.70	1.70	$\overline{3.62 = \chi^2}$
σ	1.65	1.73	
α_3	1.0	1.0	
β_{2-3}	0.6	0.6	

Table 16. Comparison of Theoretical and Actual Score Distributions for Test 2 (Difficult Items).

Score	Theoretical Frequency (F_t)	Actual Frequency (F_o)	$\frac{(F_o - F_t)^2}{F_t}$
8	8.0	8	0
7	15.0	10	1.67
6	18.8	25	2.04
5	20.5	18	.30
4	21.1	24	.40
3	20.0	22	.20
2	16.9	13	.90
1	11.3	11	.00
0	4.6	5	
M	4.21	4.21	$\overline{5.51 = \chi^2}$
σ	2.13	2.07	
α_3	0.0	-0.1	
β_{2-3}	-0.9	-0.8	

Table 17. Comparison of Theoretical and Actual Score Distributions for Test 5 (Items of Medium Difficulty).

Score	Theoretical Frequency (F_t)	Actual Frequency (F_o)	$\frac{(F_o - F_t)^2}{F_t}$
11	26.5	26	.01
10	33.2	35	.10
9	27.9	25	.30
8	20.0	22	.20
7	12.9	13	.00
6	7.8	8	.01
5	4.4	4	} .13
4	2.2	1	
3	1.0	1	
2	.3	1	
1	.1	--	
0	--	--	

M	8.85	8.86	} ~~.75~~ ²
σ	1.83	1.83	
α_3	-1.0	-1.0	
β_{2-3}	0.7	1.0	

Table 18. Comparison of Theoretical and Actual Score Distributions for Test 8 (Easy Items).

Score	Theoretical Frequency (F_t)	Actual Frequency (F_o)	$\frac{(F_o - F_t)^2}{F_t}$
9	30.2	28	.16
8	37.0	40	.24
7	29.4	28	.07
6	19.2	21	.17
5	11.0	11	0
4	5.6	8	} .09
3	2.4		
2	.8		
1	.1		
0	--		
M	7.18	7.18	.73 $\rightarrow$ σ^2
σ	1.56	1.55	
α_3	-0.9	-1.0	
β_{2-3}	0.4	1.1	

Table 19. Comparison of Theoretical and Actual Score Distributions for Test 8: (Easy Items).

Score	Theoretical Frequency (F_t)	Actual Frequency (F_o)	$\frac{(F_o - F_t)^2}{F_t}$
10	.7	1	.14
9	3.4	4	
8	8.2	6	
7	15.4	16	
6	24.5	23	.09
5	31.0	35	.52
4	25.7	21	.86
3	15.9	21	1.64
2	7.9	6	.58
1	3.0	2	
0	.7	1	
M	5.00	5.00	
σ	1.83	1.81	$3.85 \rightarrow \chi^2$
α_3	0.0	0.1	
ρ_{2-3}	-0.2	0.0	

Table 20. Comparison of Theoretical and Actual Score Distributions for Test 82 (Five Easy and Five Difficult Items).

Score	Theoretical Frequency (F_t)	Actual Frequency (F_o)	$\frac{(F_o - F_t)^2}{F_t}$
10	2.0	1	.40
9	4.4	7	
8	7.1	4	
7	9.9	15	
6	12.9	6	3.69
5	15.8	20	1.12
4	18.4	11	2.98
3	20.3	28	2.92
2	20.7	25	.89
1	17.4	11	2.35
0	7.2	8	.09
M	3.93	3.93	$18.42 \rightarrow \chi^2$
σ	2.46	2.44	
α_3	0.4	0.5	
ρ_{2-3}	-0.6	-0.6	

Table 21. Comparison of Theoretical and Actual Score Distributions for Test h (Highly Discriminating Items).

Score	Theoretical Frequency (F_t)	Actual Frequency (F_o)	$\frac{(F_o - F_t)^2}{F_t}$
10	.1	--	.36
9	1.9	2	
8	10.1	8	
7	23.4	28	
6	32.4	31	
5	30.9	30	.03
4	21.4	19	.27
3	11.0	13	.36
2	3.9	4	.01
1	.8	1	
0	.1	--	
M	5.43	5.43	1.99 χ^2
σ	1.59	1.59	
α_3	-0.2	-0.3	
$\beta_2 - 3$	-0.2	-0.4	

Table 22. Comparison of Theoretical and Actual Score Distributions for Test L (Poorly Discriminating Items).

Score	Theoretical Frequency (F_t)	Actual Frequency (F_o)	$\frac{(F_o - F_t)^2}{F_t}$
9	2.0	1	.00
8	7.6	9	
7	13.7	15	
6	19.4	18	.10
5	23.4	21	.25
4	24.2	23	.06
3	21.4	27	1.47
2	15.0	16	.07
1	7.2	6	1.00
0	1.8	--	
M	4.47	4.47	3.07 χ^2
σ	2.00	1.97	
α_3	0.1	0.1	
$\beta_2 - 3$	-0.6	-0.7	

Table 23. Comparison of Theoretical and Actual Distributions for Test r (Item Difficulties Rectangularly Distributed).

		Score on Test 5									Total
		0	1	2	3	4	5	6	7	8	
Score on Test 2	9	-	-	-	-	-	-	-	-	-	-
	8	-	-	-	-	-	-	-	0.1	0.1	0.3
	7	-	-	-	-	-	-	0.1	0.4	0.5	3 1.1
	6	-	-	-	-	-	0.3	0.5	1.1	1.1	2 3.0
	5	-	-	-	0.1	0.3	0.7	1.4	1.9	1.5	5 5.8
	4	-	-	1	0.1	0.4	0.8	1.6	2.4	2.7	10 9.8
	3	-	1	1	2	3	5	3	3	2	20 15.9
	2	-	2	2	2	4	4	6	-	1	21 25.6
	1	1	3	3	8	6	4	5	1	-	31 36.7
	0	4	5	6	10	9	5	4	1	-	44 37.9
Total		5 4.6	11 11.2	13 16.9	22 20.0	24 21.1	18 20.5	25 18.8	10 15.0	8 8.2	N=136

Table 24. Scatterdiagram for Test 2 (difficult items) and Test 5 (items of medium difficulty) showing actual frequencies (integers) and theoretical frequencies (decimals).

		Score on Test 2										Total
		0	1	2	3	4	5	6	7	8	9	
Score on Test 8	11	3 2.0	5 4.4	4 5.2	5 5.0	4 4.2	2 3.0	2 1.8	1 0.7	- 0.1	-	26 26.4
	10	7 5.3	5 8.2	10 7.6	6 5.6	4 3.5	1 1.9	- 0.8	2 0.3	-	-	35 33.2
	9	9 7.3	8 8.7	- 6.1	5 3.3	1 1.6	2 0.7	- 0.3	-	-	-	25 28.0
	8	10 7.5	6 6.8	4 3.5	2 1.5	- 0.5	- 0.1	-	-	-	-	22 20.0
	7	7 6.3	3 4.4	- 1.6	2 0.5	1 0.1	-	-	-	-	-	13 12.9
	6	4 4.5	3 2.4	1 0.7	- 0.1	-	-	-	-	-	-	8 7.8
	5	3 2.7	- 1.2	1 0.3	-	-	-	-	-	-	-	4 4.2
	4	- 1.5	1 0.5	- 0.1	-	-	-	-	-	-	-	1 2.2
	3	- 0.7	- 0.3	1 -	-	-	-	-	-	-	-	1 1.0
	2	1 0.3	-	-	-	-	-	-	-	-	-	1 0.3
	1	-	-	-	-	-	-	-	-	-	-	-
	0	-	-	-	-	-	-	-	-	-	-	-
Total		44 38.1	31 36.9	21 25.2	20 16.0	10 10.1	5 5.7	2 2.9	3 1.0	- 0.1	-	N=136

Table 25. Scatterdiagram for Test 8 (easy items) and Test 2 (difficult items) showing actual frequencies (integers) and theoretical frequencies (decimals).

		<u>Score on Test 5</u>									Total
		0	1	2	3	4	5	6	7	8	
Score on Test 8	11	-	-	-	-	1	8	8	5	4	26
		-	0.3	0.7	1.6	2.9	4.5	5.8	6.3	4.4	26.4
	10	1	-	1	5	7	6	8	3	4	35
		0.1	0.8	2.2	3.8	5.4	6.4	6.5	5.2	2.6	33.0
	9	-	-	1	7	6	3	6	2	-	25
		0.4	1.6	3.4	4.9	5.4	5.0	3.9	2.3	0.8	27.9
	8	1	3	4	8	3	1	2	-	-	22
		0.7	2.2	3.5	4.2	3.8	2.9	1.8	0.8	0.3	20.1
	7	1	3	3	1	4	-	1	-	-	13
		0.8	2.2	3.0	2.9	2.0	1.2	0.5	0.3	-	12.9
	6	-	2	2	1	3	-	-	-	-	8
		0.8	1.8	2.0	1.5	1.0	0.4	0.1	-	-	7.6
	5	2	2	-	-	-	-	-	-	-	4
		0.7	1.2	1.1	0.7	0.3	0.1	-	-	-	4.1
	4	-	1	-	-	-	-	-	-	-	1
		0.5	0.7	0.5	0.3	0.1	-	-	-	-	2.2
	3	-	-	1	-	-	-	-	-	-	1
		0.3	0.3	0.1	0.1	-	-	-	-	-	0.8
	2	-	-	1	-	-	-	-	-	-	1
		0.1	0.1	-	-	-	-	-	-	-	0.3
	1	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-
	0	-	-	-	-	-	-	-	-	-	-
		-	-	-	-	-	-	-	-	-	-
Total		5	11	13	22	24	18	25	10	8	N=136
		4.5	11.2	16.6	20.0	20.9	20.5	18.8	14.8	8.0	

Table 26. Scatterdiagram for Test 8 (easy items) and Test 5 (items of medium difficulty) showing actual frequencies (integers) and theoretical frequencies (decimals).

		Score on Test 8'										Total
		0	1	2	3	4	5	6	7	8	9	
Score on Test r	9	-	-	-	-	-	-	-	-	-	1	1
		-	-	-	-	-	-	-	0.1	0.5	1.4	2.0
	8	-	-	-	-	-	-	-	1	5	3	9
		-	-	-	-	-	-	0.1	0.7	2.4	4.2	7.5
	7	-	-	-	-	-	-	1	2	5	7	15
		-	-	-	-	-	0.1	0.5	2.0	4.9	6.1	13.7
	6	-	-	-	-	-	-	-	3	4	11	18
		-	-	-	-	0.1	0.4	1.5	3.9	7.1	6.5	19.6
	5	-	-	-	-	-	1	1	4	12	3	21
		-	-	-	-	0.3	1.1	2.9	5.7	7.8	5.4	23.1
	4	-	-	-	-	1	4	9	4	4	1	23
		-	-	-	0.1	0.7	2.0	4.2	6.5	6.8	3.7	24.1
	3	-	-	-	1	-	2	5	9	9	1	27
		-	-	0.1	0.4	1.4	2.9	4.6	5.6	4.6	1.9	21.5
	2	-	1	-	1	2	2	4	4	1	1	16
		-	-	0.3	0.7	1.6	2.7	3.4	3.3	2.2	0.7	14.8
	1	-	-	-	-	1	1	1	1	-	-	4
		-	0.1	0.3	0.7	1.2	1.5	1.5	1.2	0.5	0.1	7.2
	0	-	-	-	-	1	1	-	-	-	-	2
		-	-	0.1	0.3	0.4	0.4	0.3	0.1	0.1	-	1.8
Total		-	1	-	2	5	11	21	28	40	28	N=136
		-	0.1	0.8	2.2	5.7	11.2	19.0	29.2	37.0	30.1	

Table 27. Scatterdiagram for Test 8' (easy items) and Test r (item difficulties rectangularly distributed) showing actual frequencies (integers) and theoretical frequencies (decimals).

		Score on Test 5									Total
		0	1	2	3	4	5	6	7	8	
Score on Test 82	10	-	-	-	-	-	-	-	-	1	1
		-	-	-	-	-	-	0.1	0.3	0.3	0.7
	9	-	-	-	-	-	-	1	1	2	4
		-	-	-	-	0.1	0.3	0.7	1.1	1.2	3.4
	8	-	-	-	-	-	1	4	1	-	6
		-	-	-	0.1	0.5	1.1	1.9	2.6	1.9	8.2
	7	-	-	-	1	2	2	4	5	2	16
		-	0.1	0.3	0.8	1.8	2.9	3.8	3.7	2.0	15.4
	6	-	-	-	3	5	5	5	2	3	23
		0.1	0.5	1.4	2.7	4.2	5.2	5.2	3.7	1.5	24.5
	5	1	1	6	3	7	7	9	1	-	35
		0.4	1.6	3.5	5.4	6.3	5.8	4.5	2.4	0.8	30.9
	4	1	2	2	10	2	3	1	-	-	21
		0.8	2.9	4.8	5.6	5.0	3.7	2.0	1.0	0.3	26.0
	3	1	5	3	4	7	-	1	-	-	21
		1.2	3.0	3.8	3.4	2.4	1.4	0.5	0.1	-	15.9
	2	2	1	1	1	1	-	-	-	-	6
		1.1	2.0	2.0	1.4	0.8	0.3	0.1	-	-	7.8
	1	-	2	-	-	-	-	-	-	-	2
		0.7	1.0	0.7	0.4	0.1	-	-	-	-	2.9
	0	-	-	1	-	-	-	-	-	-	1
		0.3	0.3	0.1	-	-	-	-	-	-	0.7
Total		5	11	13	22	24	18	25	10	8	N=136
		4.6	11.4	16.6	19.9	21.4	20.5	18.9	14.8	8.0	

Table 28. Scatterdiagram for Test 82 (5 easy and 5 difficult items) and Test 5 (items of medium difficulty) showing actual frequencies (integers) and theoretical frequencies (decimals).

		Score on Test r										Total
		0	1	2	3	4	5	6	7	8	9	
S c a t t e r d i a g r a m	10	-	-	-	-	-	-	-	0.1	1	-	1 0.5
	9	-	-	-	-	0.1	0.3	0.5	1	2	1	4 3.5
	8	-	-	-	0.1	1	1.1	2	1	2	-	6 8.2
	7	-	-	0.3	1	1.8	5	4	5	1	-	16 15.4
	6	-	0.3	1	3	1	6	6	3	3	-	23 24.3
	5	0.1	0.8	4	5	10	7	5	4	-	-	35 30.7
	4	0.3	1.6	2	9	6	2	1	-	-	-	21 25.6
	3	1	2	5	7	4	1	-	1	-	-	21 15.8
	2	0.5	1	2	1	1	-	-	-	-	-	6 8.0
	1	0.4	0.8	1	1	-	-	-	-	-	-	2 3.1
	0	0.1	0.3	1	-	-	-	-	-	-	-	1 .7
Total		2 1.9	4 7.3	16 15.1	27 21.2	23 24.2	21 23.5	18 19.3	15 13.6	9 7.6	1 2.0	N=136

Table 29. Scatterdiagram for Test 82 (5 easy and 5 difficult items) and Test r (item difficulties rectangularly distributed) showing actual frequencies (integers) and theoretical frequencies (decimals).

		Score on Test L										Total		
		0	1	2	3	4	5	6	7	8	9	10		
Score on Test h	10	-	-	-	-	-	0.1	0.3	0.7	1	0.7	0.3	-	1 2.0
	9	-	-	-	-	0.1	0.4	1.0	1.5	2	1.2	0.3	-	7 4.5
	8	-	-	-	-	0.3	0.8	3	1	2.3	1.5	0.3	-	4 6.9
	7	-	-	-	0.1	0.5	3	2	8	2	1.6	0.3	-	15 9.9
	6	-	-	-	0.3	1.0	2	3	1	3.5	1.5	0.3	-	6 12.9
	5	-	-	0.1	0.5	3	5	6	4	2	1.4	0.1	-	20 15.8
	4	-	-	0.3	1.0	2	3	2	2	2	1.1	0.1	-	11 18.5
	3	-	-	1	6	2	8	7	4	-	0.7	0.1	-	28 20.3
	2	-	0.1	0.8	3	5	6	4	5	1	0.4	-	-	25 20.5
	1	-	0.3	1.4	3	4	1	1	1	-	0.1	-	-	11 17.3
	0	-	0.3	1	1	3	1	1	-	-	-	-	-	8 7.1
Total		-	1 0.7	4 3.9	13 11.0	19 21.5	30 30.9	31 32.5	28 23.3	8 10.2	2 1.8	-	N=136	

Table 30. Scatterdiagram for Test L (low validity items) and Test h (high validity items) showing actual frequencies (integers) and theoretical frequencies (decimals).

		Score on Test 5			
		0-1	2-3	4-5	6-8
Score on Test 2	4-9	$F_o = 11$		$F_o = 19$ $F_t = 19.1$ $\chi_a^2 = .00$	
		$F_t = 9.4$			
	2-3	$\chi_a^2 = .27$		$F_o = 16$ $F_t = 15.6$ $\chi_a^2 = .01$	$F_o = 26$
				$F_t = 26.1$	
	0-1	$F_o = 13$ $F_t = 14.4$ $\chi_a^2 = .14$	$F_o = 27$ $F_t = 28.7$ $\chi_a^2 = .10$	$F_o = 24$ $F_t = 22.4$ $\chi_a^2 = .11$	$\chi_a^2 = .00$

Table 31. Bivariate Grouped Frequency Table for Tests 5 and 2, Showing Observed Frequencies (F_o), Theoretical Frequencies (F_t), and $\chi_a^2 = (F_o - F_t)^2 / F_t$. ($\chi^2 = .63$).

		Score on Test 2		
		0-1	2-3	4-9
S c o r e s	10-11	$F_o = 20$ $F_t = 19.9$ $\chi_a^2 = .00$	$F_o = 25$ $F_t = 23.4$ $\chi_a^2 = .11$	$F_o = 20$
	8-9	$F_o = 33$ $F_t = 30.3$ $\chi_a^2 = .24$	$F_o = 11$ $F_t = 14.4$ $\chi_a^2 = .80$	$F_t = 19.6$
	0-7	$F_o = 27$ $F_t = 28.1$ $\chi_a^2 = .04$		$\chi_a^2 = .01$

Table 32. Bivariate Grouped Frequency Table
for Tests 2 and 8, Showing Observed
Frequencies (F_o), Theoretical Fre-
quencies (F_t), and $\chi_a^2 = (F_o - F_t)^2 / F_t$.
($\chi^2 = 1.20$).

		Score on Test 5			
		0-2	3-4	5-6	7-8
Score on Test 8	10-11	$F_o = 11$	$F_o = 13$ $F_t = 13.7$ $\chi_a^2 = .04$	$F_o = 30$ $F_t = 23.2$ $\chi_a^2 = 1.99$	$F_o = 16$ $F_t = 18.5$ $\chi_a^2 = .34$
	8-9	$F_t = 15.9$ $\chi_a^2 = 1.51$	$F_o = 24$ $F_t = 18.3$ $\chi_a^2 = 1.78$	$F_o = 15$	
	0-7	$F_o = 27$ $F_t = 25.1$ $\chi_a^2 = .14$		$F_t = 20.4$ $\chi_a^2 = 1.43$	

Table 33. Bivariate Grouped Frequency Table for Tests 5 and 8, Showing Observed Frequencies (F_o), Theoretical Frequencies (F_t), and $\chi_a^2 = (F_o - F_t)^2 / F_t$. ($\chi^2 = 7.22$).

		Score on Test 5			
		0-2	3-4	5-6	7-8
82 T e s t S c o r e	8-10	$F_o = 13$		$F_o = 11$ $F_t = 12.2$ $\chi_a^2 = .12$	
	6-7	$F_t = 16.4$	$F_o = 11$ $F_t = 9.5$ $\chi_a^2 = .24$	$F_o = 16$ $F_t = 17.1$ $\chi_a^2 = .07$	$F_o = 13$
	4-5	$\chi_a^2 = .70$	$F_o = 22$ $F_t = 22.3$ $\chi_a^2 = .00$	$F_o = 20$ $F_t = 16.0$ $\chi_a^2 = 1.00$	$F_t = 15.5$
	0-3	$F_o = 16$ $F_t = 16.2$ $\chi_a^2 = .00$	$F_o = 14$ $F_t = 10.8$ $\chi_a^2 = .95$		$\chi_a^2 = .40$

Table 34. Bivariate Grouped Frequency Table for Tests 5 and 82, Showing Observed Frequencies (F_o), Theoretical Frequencies (F_t), and $\chi_a^2 = (F_o - F_t)^2 / F_t$. ($\chi^2 = 3.49$).

		Score on Test 8'		
		0-5	6-7	8-9
r T e s t S c o r e	7-9	$F_o = 9$	$F_o = 25$ $F_t = 22.9$ $\chi_a^2 = .19$	
	5-6	$F_t = 9.6$	$F_o = 8$ $F_t = 14.0$ $\chi_a^2 = 2.57$	$F_o = 30$ $F_t = 26.8$ $\chi_a^2 = .38$
	3-4	$\chi_a^2 = .04$	$F_o = 27$ $F_t = 20.9$ $\chi_a^2 = 1.78$	$F_o = 15$ $F_t = 17.0$ $\chi_a^2 = .24$
	0-2	$F_o = 10$ $F_t = 10.3$ $\chi_a^2 = .01$	$F_o = 12$ $F_t = 13.4$ $\chi_a^2 = .15$	

Table 35. Bivariate Grouped Frequency Table for Tests 8' and r, Showing Observed Frequencies (F_o), Theoretical Frequencies (F_t), and $\chi_a^2 = (F_o - F_t)^2 / F_t$. ($\chi^2 = 5.35$).

		Score on Test r			
		0-2	3-4	5-6	7-9
82 Test Score	8-10	$F_o = 7$		$F_o = 10$ $F_t = 11.5$ $\chi^2_a = .20$	
	6-7	$F_t = 11.5$ $\chi^2_a = 1.76$		$F_o = 21$ $F_t = 17.8$ $\chi^2_a = .58$	$F_o = 17$
	4-5	$F_o = 21$ $F_t = 22.7$	$F_o = 30$ $F_t = 24.4$ $\chi^2_a = 1.28$	$F_o = 15$ $F_t = 18.3$ $\chi^2_a = .60$	$F_t = 15.2$
	0-3	$\chi^2_a = .13$	$F_o = 15$ $F_t = 14.2$ $\chi^2_a = .04$		$\chi^2_a = .21$

Table 36. Bivariate Grouped Frequency Table for Tests r and 82, Showing Observed Frequencies (F_o), Theoretical Frequencies (F_t), and $\chi^2_a = (F_o - F_t)^2 / F_t$. ($\chi^2 = 4.80$).

		Score on Test L			
		0-3	4-5	6-7	8-10
h T e s t s o n S c o r e	6-10		$F_o = 28$ $F_t = 28.4$ $\chi_a^2 = .01$		$F_o = 9$ $F_t = 10.7$
	4-5	$F_o = 18$	$F_o = 13$ $F_t = 12.8$ $\chi_a^2 = .00$	$F_o = 14$ $F_t = 16.9$ $\chi_a^2 = .50$	$\chi_a^2 = .27$
	2-3	$F_t = 15.3$	$F_o = 21$ $F_t = 20.1$ $\chi_a^2 = .04$	$F_o = 24$ $F_t = 19.4$	
	0-1	$\chi_a^2 = .48$	$F_o = 9$ $F_t = 12.3$ $\chi_a^2 = .88$	$\chi_a^2 = 1.09$	

Table 37. Bivariate Grouped Frequency Table for Tests L and h, Showing Observed Frequencies (F_o), Theoretical Frequencies (F_t), and $\chi_a^2 = (F_o - F_t)^2 / F_t$. ($\chi^2 = 3.27$).

h_i = a measure of the difficulty of item i , defined in terms of p_i
by the relation $p_i = A(h_i)$.

i = subscript indicating the i -th item; $i = 1, 2, \dots, n$. In M_i, σ_i, r_{ic} ,
etc. i is used to stand for x_i .

j = subscript indicating the j -th item; $j = 1, 2, \dots, n$. In M_j, σ_j, r_{jc} ,
etc. j is used to stand for x_j .

$$K_i = \sqrt{1 - R_i^2}.$$

k = the number of choices in a multiple choice item.

M_y = the mean of any variable, y .

$M_{u.v}$ = the conditional mean of any variable, u , for a fixed value of
another variable, v ; the regression of u on v .

$M(c) = M_p$ = the mean of the values of P_i for any given value of c .

n = number of items in the test.

$N(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$ = the normal probability density function for a
standardized variable y ; the normal curve ordinate at any point,
 y .

$N(y; M, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left[-\frac{1}{2\sigma^2} (y - M)^2 \right]$ = the normal frequency function
of any variable, y , with mean M and standard deviation σ .

$N_2(u, v; r) = \frac{1}{2\pi\sqrt{1-r^2}} \exp \left[-\frac{u^2 + v^2 - 2ruv}{2(1-r^2)} \right]$ = the normal bivariate
frequency function for any pair of standardized variables, u
and v , whose correlation is r .

p_i = the "item difficulty"; the proportion of all examinees answering
item i correctly.

$P_i = A(g_i)$ = the probability that an examinee at a given level of ability will answer item $\underline{i}$ correctly (P_i is a function of $\underline{c}$).

$q_i = 1 - p_i$.

$Q_i = 1 - P_i$.

r_{uv} = the correlation between any variables, $\underline{u}$ and $\underline{v}$ (except where otherwise indicated).

r_{ss} = the parallel-form reliability of the test score.

r_{ic} = the product moment (point biserial) correlation between $\underline{c}$ and x_i .

R_i = biserial correlation of x_i and $\underline{c}$; loading of item $\underline{i}$ on the common factor (c), as calculated from the matrix of tetrachoric item intercorrelations.

r'_{ij} = tetrachoric correlation between items $\underline{i}$ and $\underline{j}$.

r_{ij} = the product moment (fourfold point) correlation between x_i and x_j .

$s = \sum_{i=1}^n x_i$ = the test score.

$t = \lim_{n \rightarrow \infty} z$ = the "relative true score"; the proportion of correct answers on an infinitely long test.

u = any variable.

v = any variable.

x_i = the score on item $\underline{i}$; $x_i = 0$ or 1 .

y = any variable.

$z = \frac{s}{N}$ = the "relative score"; the proportion of items answered correctly.

γ_{sc} = the curvilinear correlation or correlation ratio of test score on ability.

π_s, π_{n-s} (see explanation of Equation B1.2)

$\sigma_1 = \sqrt{p_1 q_1}$ = the standard deviation of x_1 .

σ_y = standard deviation of any variable, y .

Σ^* (see explanation of Equation B1.2)