Recommended citation:
Lord, F. (1952). *A Theory of Test Scores* (Psychometric Monograph No. 7). Richmond, VA: Psychometric Corporation. Retrieved from http://www.psychometrika.org/journal/online/MN07.pdf

A THEORY OF TEST SCORES

By

FREDERIC LORD

EDUCATIONAL TESTING SERVICE

PREFACE

The present monograph is concerned with the relationship between the obtained test score and the underlying trait or ability involved in taking the test. It is shown that this relationship is a function of the difficulties and intercorrelations of the items of which the test is composed; the relationship is not a simple one. The resultant theory of mental test scores is more appropriate and more powerful in the area for which it is intended than is a direct application of the classical theory of errors, starting with the broad assumption that test score and "true score" differ by normally distributed, independent errors of measurement. The present theory of test scores starts with assumptions designed to fit certain testing situations, and proceeds to investigate the shapes of the frequency distributions of test scores, of true scores, and of errors of measurement, and further, the relation of these variables to the "ability" involved in taking the test. The conclusions reached do not in general contradict the basic formulas already firmly established in mental test theory, such as the Spearman-Brown formula and the formula for correction for attenuation. A number of new conclusions are reached, however. Some of them are at variance with certain commonly held conceptions; for example, it is found that the regression of test score and of true score on "ability" is in general necessarily curvilinear and that the errors of measurement have a binomial distribution that is not independent of true score.

The theory presented here is directly applicable only to tests composed of free-response items. The theoretical extension to cover items that may be answered correctly by guessing is straightforward mathematically but seems to require assumptions that may be poorly fulfilled in actual practice. A number of conclusions of general validity nevertheless can be reached. For the sake of brevity and clarity, this extension is not presented here, but the broad outlines of such an extension will suggest themselves to the reader.

The main function of the present monograph is to outline the framework of a logical structure (mathematical model) that, it is hoped, will serve as a basis for further development of our understanding of test scores and their relation to the trait to be meas-

ured. It is believed that the test technician will see much here that is relevant to the practical problems of selecting the items for building a test to be used for a specific purpose. In particular, he should find of interest the discussion on the discriminating power of the test at various ability levels, although much more remains to be done on this important problem. After the test has been built and administered, much will also be found in the present theory that is relevant to the interpretation of the scores actually obtained.

It will be helpful if the reader has some familiarity with integral calculus and with the mathematics of frequency distributions. The latter subject is covered by many texts, such as that by S. S. Wilks, to which specific reference will be made, or such as the recent "Introduction to the Theory of Statistics" by A. M. Mood (McGraw-Hill, 1950).

The present monograph, except for numerous revisions of a fairly superficial nature, was presented as a doctoral dissertation at Princeton University in March, 1951. The writer wishes to express to his adviser, Professor Harold Gulliksen, and to Professor Ledyard R. Tucker his great appreciation for their interest, advice, and criticism. Several of the developments presented here have arisen from Professor Tucker's suggestions. Thanks are also due to Professor John Tukey for his suggestions on a recent draft of the manuscript.

The raw data used as a basis for empirical verification of the theoretical results were kindly loaned by Dr. Lynnette B. Plumlee, who had painstakingly gathered them for other research purposes. Acknowledgment is due Dr. Dorothy C. Adkins for her valuable editorial work; also to Mrs. Ruth Blackman, Miss Henrietta Gallagher, Miss Lorraine Luther, Miss Braxton Preston, and Mrs. Mary Evelyn Runyon of the Educational Testing Service for their excellent work in connection with the computing, typing, and proofreading.

Support for a large part of this work was provided by the Educational Testing Service. Special thanks are due Dr. Henry Chauncey for his generous interest.

FREDERIC M. LORD

February, 1952

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PART 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 shall 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; hence its magnitude, in terms of whatever metric may be chosen, 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 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 skewed positively and the other negatively, 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 ones in the latter. 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.

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 then be completely determined as a function of the usual item statistics. If this invariant metric is accepted 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, the result is that all the properties of the test score in relation to the underlying ability will thus have been expressed as functions of the usual item statistics.

In this way answers to such questions as the following can be attempted: Under what circumstances, if any, can platykurtic, rectangular, or U-shaped distributions of test scores be expected? Do equal units of test score correspond to equal units of ability, as here defined? Does the standard error of measurement vary with the ability level of the examinees tested and, if so, in what fashion? 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 the items that compose the test be selected in such a way as to achieve any desired result?

The present monograph 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.

Lawley (20, 21) has published a number of important theoretical developments, similar to certain of those to be given here, for 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 (and also the present writer) assumes the probability that an examinee will answer an item correctly to be a normal-ogive function of the examinee's ability (see Equation 2). More recently, Carroll (2) has offered a similar theoretical development based on the less usual assumption that this probability is a linear function of ability within certain limits and is equal to 0 or 1 outside these limits. 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.

The normal-ogive assumption that will be used here previously has been applied with considerable success by Tucker in a theoretical study (36) of the effect on test validity of changes in the intercorrelation of the items, and in a second study (37) to predict how different groups of examinees will perform on an item. Brogden (1) has made similar assumptions in a theoretical and numerical study of the relation of item intercorrelation and of item difficulty distribution to test validity and reliability.* Cronbach and Warrington (4) have recently made use of similar assumptions to infer valuable generalizations about the discriminating power of certain multiple-choice tests at various levels of test score.

* The same assumptions are used in Lord, F. M. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 1952, 17.

PART II

THEORETICAL TREATMENT

A. RESTRICTIONS, ASSUMPTIONS, AND DEFINITIONS

A1. *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 $(i = 1, 2, \ldots, n)$ 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.

Consideration will be restricted further to the case where the test score (s) is the sum of the scores on the n items of which the test is composed:

$$s = \sum^{n} x_i. \tag{1}$$

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

A2. *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, this relationship may be stated more explicitly:

$$P_i = \int_{-\infty}^{\frac{c - a_i}{b_i}} N(y)\,dy, \tag{2}$$

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,

$$N(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}. \tag{3}$$

The notation $A(y_o)$ will be used to indicate the area of the standardized normal curve lying above the given point, y_o. Thus, because of the symmetry of the normal curve, we may write

$$P_i = A\left(\frac{a_i - c}{b_i}\right). \tag{4}$$

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 (22) and others have called the *trace line* of the item.

The relationship expressed by (2) has been widely used in psychophysics in other connections and in recent years has been used to represent the probability of answering a test item correctly by Guilford (11), Richardson (32), Mosier (26, 27), Ferguson (7), Lawley (20, 21), Lorr (24), Tucker (36), Cronbach and Warrington (4), 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.

It should be noted that the relationship given by (2) is reasonable only in the case of free-response items that cannot be answered correctly by guessing. An item requiring the examinee to add a column of figures, for example, would probably meet these requirements. If the examinee is not able to add correctly, guessing will be of little help to him in obtaining the correct sum. The present theory may be amplified to cover multiple-choice items, but this extension will not be treated here.

In order to use Equation (2), the values of a_i and b_i for each item must first be determined. If no further assumptions are made, this presumably could be achieved 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 supposedly 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, 11, 173; also Finney (9), who presents 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 a further restriction on the data to be considered has been discussed.

A3. *Restriction on the Population of Examinees*

First let us limit consideration to a sample of examinees so large that sampling problems need not enter into the discussion. Such problems, while of practical importance, 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. For many developments, it will be more profitable, however, to restrict attention to the case where ability is normally distributed in the group of examinees tested. A normal distribution of ability will be assumed when not otherwise specified. Many of the formulas to be obtained, however, 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.

In any case, the scale of measurement for ability may be chosen, without loss of generality, so that the mean (M) and standard deviation (σ) of c are, respectively, 0 and 1:

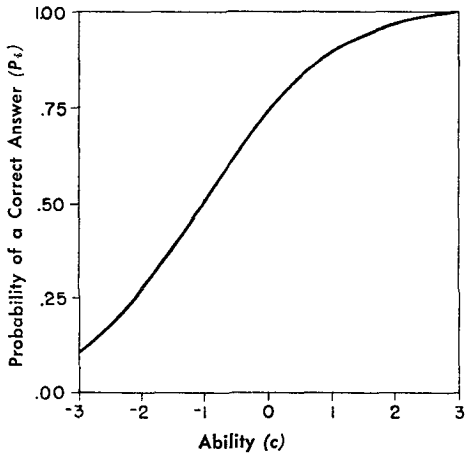
$$M_c = 0, \sigma_c = 1. \tag{5}$$

The assumption of a normal distribution of ability in the group tested will simplify the determination of the values of a_i and b_i for each item. Tucker (36, 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

$$R_i = \frac{1}{\sqrt{1 + b_i^2}}, \tag{6}$$

$$p_i = A(h_i), \tag{7}$$

where $h_i = a_i R_i$.



Item Characteristic Curve When $h_i = -.562$ and $R_i = .531$

FIGURE 1

It will be assumed henceforth, unless otherwise specified, that $R_i^2 < 1$. We may then write

$$K_i = \sqrt{1 - R_i^2}, \tag{8}$$

$$g_i = \frac{h_i - R_i c}{K_i}. \tag{9}$$

Let us use (6) and (7) to eliminate a_i and b_i from the equation for the item characteristic curve (4) so as to obtain the following formula, which will be most convenient for our further development:

$$P_i = A(g_i). \tag{10}$$

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 (10), represents

an item for which $h_i = -.562$ and $R_i = .531$ (item 44 of Part III). The curve is asymptotic to $P_i = 0$ and $P_i = 1$.

The numerical values of item characteristic curves may be determined approximately in practice from appropriate empirical data without the use of probit analysis methods. The actual numerical value of h_i can readily be obtained for any given set of data from the item difficulty, by means of (7). An approximation to the value of R_i perhaps can be obtained from the biserial correlation of the item with score on a very long test of the ability under consideration, after this biserial correlation has been corrected for attenuation due to the unreliability of the test. A more accurate method of determining R_i will be discussed in the following section.

A4. *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 consideration will be limited 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 (38) for a discussion of the reasonableness of such a restriction]. From this restriction it follows that, for fixed c, P_i is independent of P_j, where i and j represent any two different test items.

Since c is the common factor in the matrix of tetrachoric item intercorrelations, 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.

A5. *Alternative Assumptions and Restrictions*

Before proceeding with our development, it will be of interest to point out, as Brogden (1, 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 $2 \times 2 \times \ldots \times 2$ table in n dimensions containing 2^n cells. The equivalent set of assumptions is 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 conforms to these two requirements. The first of these two requirements cannot be contradicted by the data unless $n \geqslant 3$, the second unless $n \geqslant 4$.

B. THE BIVARIATE DISTRIBUTION OF TEST SCORE AND ABILITY

B1. *The Frequency Distribution of Scores for Examinees at a Given Level of Ability*

Since c is the only common factor of the tetrachoric item intercorrelations, it follows, as already has been pointed out, that P_i and P_j $(i \neq j)$ are independent when c is fixed. It is thus seen, for example, that the probability of success on both items i and j is P_iP_j; that the probability of success on one and failure on one is $P_iQ_j + Q_iP_j$, where $Q_i = 1 - P_i$; and finally that the probability of failure on both items is Q_iQ_j. Generalizing this result to n items, we find that the probability of success on s items out of n, for examinees at a given level of ability, is given by the sum of the appropriate terms in the expansion of the product

$$\prod_{i=1}^{n}(Q_i + P_i). \tag{11}$$

(It must be borne in mind here that P_i and Q_i are functions of c, as shown by Equations 10 and 9.) Since the score is the number of items answered correctly, the terms of the expansion of (11) represent the distribution of test scores (s) at a given level of ability. If we denote this conditional distribution of s by $f_{s.c}$, we obtain by expanding (11) the result that

$$f_{s.c} = \Sigma^* \Pi_s P_i \Pi_{n-s} Q_i \quad (s = 0, 1, \ldots, n), \tag{12}$$

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

possible products. If all items are equivalent, so that all P_i are equal, (11) reduces to the familiar binomial expression, and (12) becomes, upon dropping subscripts,

$$f_{s.c} = \binom{n}{s} P^s Q^{n-s}. \tag{13}$$

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 happen to be under consideration. Equations (11), (12), and (13) 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.

B2. *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 (39, 17) the product of the conditional distribution of (12) and the marginal distribution of c:

$$f_{cs} = N(c) f_{s.c} \quad (s = 0, 1, \ldots, n), \tag{14}$$

where $f_{s.c}$ is the expression given in (12). 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} can be calculated from actual item analysis data.

If for a moment the requirement that the distribution of c shall be normal in the group of examinees under consideration be dropped and the distribution be allowed to take any arbitrary form, f_c, the same line of argument that already has been used leads to a general formula for the bivariate distribution of ability and test score, as follows:

$$f_{cs} = f_c \Sigma^* \Pi_s P_i \Pi_{n-s} Q_i \quad (s = 0, 1, \ldots, n). \tag{15}$$

The values of R_i necessary for computing the expression under the summation sign in (15) 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 7, 8, and 9, at the end of this monograph, which provide numerical examples of the bivariate distribution of test score and ability.

B3. *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 ΣP_i. In other terms, the average score of examinees at a given level of ability is

$$M_{s.c} = \Sigma P_i. \tag{16}$$

Equation (16) is the usual expression for the mean of the distribution given by (12) (17, Vol. I, 122). Since P_i is a function of c, (16) is the equation for the regression of test score on ability. Like the conditional distribution of s in (12), 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 deduced previously by Brogden (1, 207).

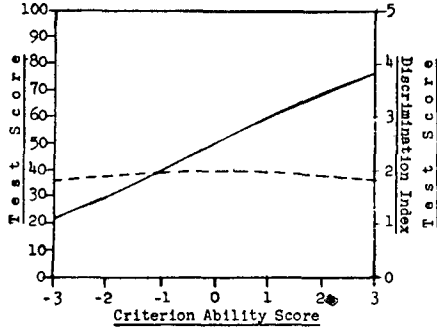
Obviously the regression curve must be the sum of the n item characteristic curves—in this case, of n normal ogives. If all the items have equal difficulties and equal intercorrelations, we have, dropping the subscript i, $M_{s.c} = nP$, and the regression curve is directly proportional to the item characteristic 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. This statement remains true no matter how ability is defined or what metric is used to measure it, and no matter what general form is assumed for the item characteristic curve. Since the regression curve must be the sum of the item characteristic curves, it is seen 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.

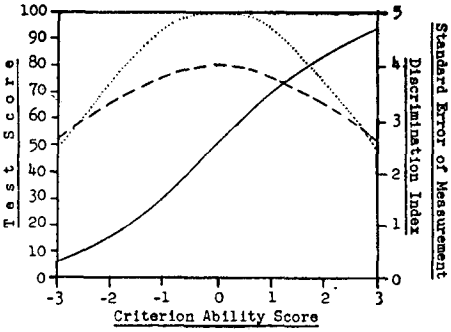
Any given group of examinees may, to be sure, have a limited range of ability within which the regression might appear to be approximately rectilinear. A derivation indicating the extent to which the regression curve will appear to be rectilinear in any given group of examinees will be presented in Section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

B4. *The Regression of Ability on Test Score*

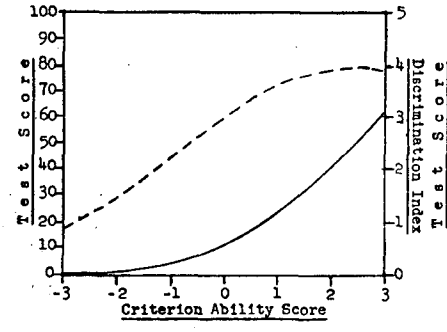
The other regression curve—the regression of ability on test score—can be written down in the form of a definite integral if



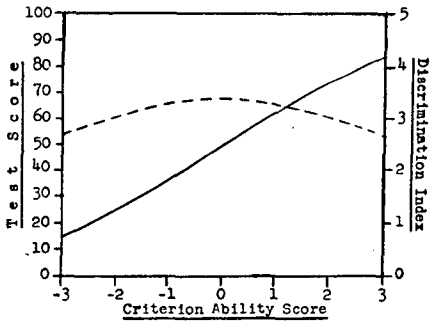
Test 1: All items of 50% difficulty; all item tetrachoric intercorrelations = .06; test reliability = .80.



Test 2: All items of 50% difficulty; all item tetrachoric intercorrelations = .20; test reliability = .94.



Test 3: All items of 14% difficulty; all item tetrachoric intercorrelations = .20; test reliability = .91.



Test 4: Item difficulties rectangularly distributed; all item tetrachoric intercorrelations = .20 (reliability not computed).

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

FIGURE 2

desired. The integration cannot in general 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 change in 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.

As an approximation to the regression of ability on test score for the particular case where all items are equivalent, one might possibly use the modal value of c for fixed s, as found by taking the partial derivative of f_{cs} with respect to c and setting the result equal to zero. Now,

$$\frac{\partial P_i}{\partial c} = \frac{R_i}{K_i} N(g_i). \tag{17}$$

Dropping subscripts, it is found from (14), (13), and (17),

$$\frac{\partial f_{cs}}{\partial c} = \binom{n}{s} P^{s-1} Q^{n-s-1} N(c) \left[\frac{R}{K} (s - nP) N(g) - cPQ \right]. \tag{18}$$

Setting this derivative equal to zero, we see that the modal value of c for fixed s is the value of c that satisfies the equation

$$KcPQ = R(s - nP)N(g). \tag{19}$$

This equation may be solved by successive approximation methods. It may be noted that the solution constitutes a maximum likelihood estimate of an examinee's ability score (c) from his obtained test score (s) when it is given that c is normally distributed in the group tested.

B5. *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 (17, Vol. I, 122)

$$\sigma_{s.c} = \sqrt{\Sigma P_i Q_i} = \sqrt{n(M_P M_Q - \sigma_P{}^2)}, \tag{20}$$

where

$$M_P = \frac{1}{n} \Sigma P_i, \quad M_Q = \frac{1}{n} \Sigma Q_i, \tag{21}$$

and

$$\sigma_P{}^2 = \frac{1}{n} \Sigma P_i{}^2 - M_P{}^2. \tag{22}$$

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 (20) 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 always exist individuals whose ability is so low that the test would not be discriminating for them, and other individuals whose ability is so high that the test would likewise not be 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 monograph.

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, 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 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 will be denoted by z, where

$$z = \frac{s}{n}. \tag{23}$$

When n becomes infinite, the relative score (z) becomes the "relative true score" which will be denoted by t, i.e.,

$$t = \lim_{n=\infty} z = \lim_{n=\infty} \frac{s}{n}. \tag{24}$$

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 it is desired to consider 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.

Since the standard deviation of z will be $1/n$ times the standard deviation of s, we see from (20) that the standard error of measurement for relative score at a given level of ability is

$$\sigma_{z.c} = \sqrt{\frac{1}{n}(M_P M_Q - \sigma_P^2)}. \tag{25}$$

Letting $n = \infty$ in (25), it is seen that $\sigma_{t.c} = 0$. This result has been pointed out by Brogden (1) and others. Thus the important conclusion is reached 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 c, then the transformed true scores will be identical with c. In particular, if c is normally distributed in the group tested, the value of c for each examinee will be identically equal to his normalized true score. Although in practice the examinee's true score is not available 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 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 c is seen in the fact that, except for errors of measurement, the same value of 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:*

$$\sigma_{s.c} = \sigma_{s.t}. \tag{26}$$

The standard deviation of c ($\sigma_{c.s}$) for individuals having a given test score can be expressed in terms of definite integrals derived from (14). The resulting expressions have so far been too intractable to be of interest here, however. The standard error of measurement for examinees *at a specified level of actual test score,*

which recently has been investigated by Mollenkopf (25, 13, 115-126), is a still different statistic and should not be confused with $\sigma_{c.s}$, $\sigma_{s.c}$, or $\sigma_{s.t}$, as discussed here.

B6. *The Product-Moment Correlation Between Test Score and Ability*

The product-moment correlation between test score and ability may be obtained from Equation (14) by the usual method of integration. The same result is commonly obtained much more easily, however, from the usual formula for the correlation between sums. The result is:

$$r_{cs} = r_{c(\Sigma x_i)} = \frac{\Sigma \sigma_i r_{ic}}{\sigma_s}. \tag{27}$$

Here σ_i, the standard deviation of x_i, may be calculated from the item difficulties by the usual formula:

$$\sigma_i = \sqrt{p_i q_i}\,; \tag{28}$$

r_{ic}, the product-moment or point-biserial (16, 18) correlation between c and x_i, may be calculated from the biserial correlation R_i by means of the usual equation relating biserial and point-biserial correlation coefficients:

$$r_{ic} = \frac{N(h_i)}{\sqrt{p_i q_i}} R_i\,; \tag{29}$$

and σ_s, the standard deviation of the test score, may be calculated from the usual formula for the standard deviation of sums,

$$\sigma_s = \sigma_{\Sigma x_i} = \sqrt{\Sigma\Sigma \sigma_i \sigma_j r_{ij}}\,, \tag{30}$$

where r_{ij} is the product-moment (fourfold-point) correlation between x_i and x_j. If $i = j$, r_{ij} is here taken to be 1; otherwise the value of r_{ij} may be calculated from the usual fourfold table or from r_{ij}', the tetrachoric correlation between x_i and x_j, by means of the standard formula (6, 124), which in our notation is

$$r_{ij} = \frac{A_2(h_i, h_j; r_{ij}') - p_i p_j}{\sqrt{p_i q_i p_j q_j}}, \tag{31}$$

where

$$A_2(h_i, h_j; r_{ij}') = \int_{h_i}^{\infty}\int_{h_j}^{\infty} N_2(u, v; r_{ij}')\,dv\,du, \tag{32}$$

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 $A_2(u_o, v_o; r)$ are tabled by Pearson (29, Vol. II).

Using Equations (27) through (30), the correlation of test score with ability may be written

$$r_{cs} = \frac{\Sigma R_i N(h_i)}{\sqrt{\Sigma\Sigma\sigma_i\sigma_j r_{ij}}} . \tag{33}$$

B7. *The Curvilinear Correlation of Test Score on Criterion Score; the Test Reliability*

The correlation ratio of test score on ability is, of course, 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, 280)

$$\eta_{sc} = \sqrt{1 - \frac{E(\sigma_{s.c}^2)}{\sigma_s^2}} , \tag{34}$$

where $E(\sigma_{s.c}^2)$ denotes the expected or average value of $\sigma_{s.c}^2$ for the total group of examinees. Now, $\sigma_{s.c} = \sigma_{s.t}$ (26); 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.$_{\text{meas.}}$). It is well known that

$$\text{S.E.}_{\text{meas.}} = \sigma_s \sqrt{1 - r_{ss}} , \tag{35}$$

where r_{ss} is the test reliability. Substituting (35) in (34) the interesting result follows that the curvilinear correlation of test score on ability is equal to the square root of the test reliability:

$$\eta_{sc} = \sqrt{r_{ss}} . \tag{36}$$

This result is a consequence of the well-known fact that $r_{st} = \sqrt{r_{ss}}$.

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

$$r_{ss} = r_{(\Sigma x_i)(\Sigma X_j)} = \frac{\Sigma\Sigma_{i\ j}\sigma_{x_i}\sigma_{X_j}r_{x_iX_j}}{\sigma_s^2} . \tag{37}$$

Equation (37) is the same as the basic equation from which Kuder and Richardson (19, Equation 2) derive their reliability formula.

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

$$\sigma_{X_j} = \sigma_j \,, \tag{38}$$

the observed variance of the j-th item (see Equation 28); and

$$r_{x_iX_j} = r_{ij} \; (i \neq j), \tag{39}$$

the observed fourfold-point correlation (see Equation 31). The value of $r_{x_iX_i}$ for our purposes may be determined from the requirement that the corresponding tetrachoric correlation, $r'_{x_iX_i}$, shall be such that the correlation matrix whose elements are $r'_{ij} = r'_{x_iX_j}$ shall have unit rank. We see from the definition of R_i in Section A4 that

$$r'_{x_iX_i} = R_i^2; \tag{40}$$

hence, by (28) and (31),

$$r_{x_iX_i} = \frac{A_2(h_i,h_i;R_i^2) - p_i^2}{\sigma_i^2}. \tag{41}$$

From Equations (30) and (36) through (41), we find the result that

$$r_{ss} = \eta_{sc}^2 = \frac{\sum_i A_2(h_i,h_i;R_i^2) - \sum_i p_i^2 + \sum_{i\neq j}\sum \sigma_i\sigma_j r_{ij}}{\sigma_s^2}$$

$$= 1 - \frac{\sum_i p_i - \sum_i A_2(h_i,h_i;R_i^2)}{\sigma_s^2}. \tag{42}$$

The expression given for the squared curvilinear correlation coefficient in (42) may be verified, if desired, by evaluating by direct integration the expected value (39, 29) of $\sigma^2_{s.c}$. Using (20), it will be found that

$$E(\sigma^2_{s.c}) = \int_{-\infty}^{\infty} (\Sigma P_iQ_i)N(c)\,dc = \Sigma p_i - \Sigma A_2(h_i,h_i;R_i^2). \tag{43}$$

Substitution of this result in (34) yields the same expression for the curvilinear correlation coefficient as that given by (42).

An examination of (42) shows that any increase in R_i^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 (36) and Brogden's 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 behavior of the product-moment correlation results in part from the fact that as the item intercorrelations increase, the regression becomes more and more curvilinear, 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.

No simple analytic results for the other curvilinear correlation (η_{cs})—that of ability on test score—have been obtained to date in view of the rather intractable expression obtained for $\sigma_{c.s}$. Numerical values of η_{cs} for a number of selected hypothetical tests have been computed by Cronbach and Warrington by means of numerical integration. Their results should throw considerable light on the properties of this statistic.

B8. *The Magnitude of the Practical Effect of the Curvilinear Regression*

As noted in Section B3, 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 a formula for the product-moment correlation for the case of equivalent items will be obtained. From (30), dropping the subscripts i and j and remembering that $r_{ii} = 1$, is obtained

$$\sigma_s = \sqrt{npq + (n^2 - n)pqr} = \sqrt{[1 + (n-1)r]npq}. \qquad (44)$$

From (27), (28), and (44) the product-moment correlation is found to be

$$r_{cs} = \frac{r_c\sqrt{n}}{\sqrt{1 + (n-1)r}}. \qquad (45)$$

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

$$R_i^2 = r'. \tag{46}$$

For equivalent items, dropping the subscripts i and j, Equation (31) becomes

$$r = \frac{A_2(h,h;r') - p^2}{pq}. \tag{47}$$

Replacing R_i^2 in (43) by r' and eliminating $A_2(h, h; r')$ from (43) by using (47), we obtain

$$E(\sigma_{s.c}^2) = npq(1-r). \tag{48}$$

From (34), (44), and (48) is thus obtained the result that for equivalent items

$$\eta_{sc}^2 = 1 - \frac{1-r}{1+(n-1)r} = \frac{nr}{1+(n-1)r}. \tag{49}$$

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, 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 (45) and (49) follows the very simple result that

$$\frac{\eta_{sc}^2}{r_{cs}^2} = \frac{r}{r_c^2}. \tag{50}$$

By (29) and (46),

$$\frac{\eta_{sc}^2}{r_{cs}^2} = \frac{pqr}{R^2N^2(h)} = \frac{pqr}{r'N^2(h)}. \tag{51}$$

In order to obtain a clearer idea of the magnitude of this ratio, it will be helpful to expand 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 (30, 369). For equivalent items, this series is

$$\frac{A_2(h,h;r') - p^2}{N^2(h)} = r' + \frac{1}{2}h^2r'^2 + \frac{1}{6}(h^2-1)^2r'^3$$

$$+ \frac{1}{24}h^2(h^2-3)^2r'^4 + \frac{1}{120}(h^4-6h^2+3)^2r'^5 + \ldots. \tag{52}$$

By (47), the left side of (52) is

$$\frac{A_2(h, h; r') - p^2}{N^2(h)} = \frac{pqr}{N^2(h)}. \tag{53}$$

From (51), (52), and (53) is obtained

$$\frac{\eta_{sc}^2}{r_{cs}^2} = 1 + \frac{1}{2}h^2r' + \frac{1}{6}(h^2 - 1)^2r'^2 + \frac{1}{24}h^2(h^2 - 3)^2r'^3 + \frac{1}{120}(h^4 - 6h^2 + 3)^2r'^4 + \ldots . \tag{54}$$

If all items are of 50 per cent difficulty, $h = 0$, and it follows that

$$\frac{\eta_{sc}^2}{r_{cs}^2} = 1 + \frac{1}{6}r'^2 + \frac{3}{40}r'^4 + \ldots . \tag{55}$$

This last equation shows that when all items are of 50 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 (54), on the other hand, shows that whenever the item difficulties differ by a considerable extent from 50 per cent,—when $h \geqslant 1$, 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 illustrations relating to these conclusions.

C. THE DISCRIMINATING POWER OF THE TEST AT A GIVEN LEVEL OF ABILITY

C1. *Derivation of an Index of Discriminating Power*

If a test is to be used for selecting those individuals having the greatest amount of the ability measured by the test, a test is desired 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 more 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 discrimi-

nation over the entire group of examinees. The standard error of measurement at a given level of ability ($\sigma_{s.c}$) is usually thought of as a direct measure of the discriminating power of the test at that level; but we have already shown (at the beginning of Section B5) that for a given test the standard error of measurement at various levels of ability tends to be inversely related to the test's discriminating power.

The standard deviation of ability at a given level of test score ($\sigma_{c.s}$) would provide a good indication of discriminating power in certain circumstances—for example, when a counsellor is considering the test score of a single individual for guidance purposes. A refinement of this approach would be to use standard methods of statistical estimation and set up for each test score a corresponding confidence interval for ability, within which the true ability score could be assumed to lie. The length or other properties of such confidence intervals could be taken as a measure of the discriminating power of the test for examinees at different levels of test score.

In other circumstances, different indices of discrimination would be required. For example, if we are trying to select the 40 examinees having the highest ability in a group of 100 examinees, we might ask what proportion of the 40 best examinees on ability would be found among the 40 examinees selected as having the highest test scores. This proportion would serve fairly well as an index of discriminating power for certain purposes. Still other indices will suggest themselves for other purposes. In particular, Cronbach and Warrington (4) have devised at least two such indices and drawn important conclusions as to the relation between test discriminating power, as measured by these indices, and the composition of the test, described in terms of the item difficulties and intercorrelations.

The discrimination indices discussed in the two preceding paragraphs relate to the discriminating power of the test *at a given level of test score,* as distinguished from the discriminating power *at a given level of ability.* This distinction is strictly analogous to the usual distinction between the standard error of a true score and the standard error of measurement, respectively. The former type of index has great practical importance because in actual practice we know the examinees' test scores exactly, but we know their ability scores or true scores only approximately. Actual cutting scores must be in terms of test scores, not of ability scores.

On the other hand, any good index of the discriminating power of a test at a given level of test score must be based on some as-

sumption or knowledge about the distribution of ability in the group tested. If the entire group of examinees tested has a high average level of ability, for example, the frequency distribution of the ability levels of those examinees who obtain any given test score will be different than if the entire group of examinees tested has a low average ability level. The discriminating power of a test at a given level of test score therefore is always a function of the nature of the group tested and cannot be invariant from group to group. If a certain test is intended to be administered to a wide variety of groups whose characteristics cannot accurately be predicted in advance, it will be helpful in describing the actual or the desirable properties of the test to consider some index of its discriminating power *at a given ability level*—an index that will remain invariant no matter to what group the test is administered. An attempt will here be made to develop such an index.

A numerical example may 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 examinees at the $c = -5$ level and examinees at the $c = -3$ level. Clearly any measure of discriminating power at a given ability level 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 sufficiently 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

$$D' = \frac{|M_{s.c_0} - M_{s.c_1}|}{\overline{\sigma^*}}, \tag{56}$$

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 $c_1 - c_0$. Interest is not 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

$$\lim_{c_1 \doteq c} \frac{|M_{s.c_1} - M_{s.c}|}{\sigma^*(c_1 - c)} = \lim \frac{1}{\sigma^*} \cdot \lim \frac{|M_{s.c_1} - M_{s.c}|}{c_1 - c} = \frac{1}{\sigma_{s.c}} \frac{\partial M_{s.c}}{\partial c}. \tag{57}$$

Obtaining the indicated derivative from (16) and (17), using the expression for $\sigma_{s.c}$ given by (20), and denoting the rate by D, we have

$$D = \frac{1}{\sqrt{\Sigma P_i Q_i}} \Sigma_i \frac{R_i}{K_i} N(g_i). \tag{58}$$

Both P_i and g_i are functions of c; so D is likewise a function of c. D will be used as the measure of the discriminating power of the test at a specified level of ability (at a specified value of c, not of s).

It is seen that 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 it will be assumed that the slope of the regression will always be non-negative. 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_i are available for any group of items meeting the assumptions made, the value of D can be calculated from Equation (58) without excessive difficulty.

It may be noted that 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 (20) 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. If in any given group of examinees there are only a few individuals spread out thinly over this range of ability, however, 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 (58) 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.

C2. *The Conditions for Maximum Discrimination*

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

$$\frac{dN_i}{dg_i} = -g_i N_i, \tag{59}$$

$$\frac{dP_i}{dg_i} = -N_i, \tag{60}$$

$$\frac{d \log D}{dg_i} = -\frac{g_i N_i}{\Sigma N_i} + \frac{(1-2P_i)N_i}{2\Sigma P_i Q_i}. \tag{61}$$

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

$$\frac{(1-2P_i)}{g_i} = \frac{2\Sigma P_i Q_i}{\Sigma N_i} \quad (i = 1, 2, \ldots, n). \tag{62}$$

Since the expression on the right is the same for all n equations, a condition for a maximum of D is that

$$\frac{1-2P_1}{g_1} = \frac{1-2P_2}{g_2} = \ldots = \frac{1-2P_n}{g_n}. \tag{63}$$

Now the absolute value of g_i is a single-valued function of $(1-2P_i)/g_i$. Consequently (63) requires for a maximum of D that all g's be equal in absolute value:

$$|g_1| = |g_2| = \ldots = |g_n|. \tag{64}$$

Under the conditions of (64), $P_1Q_1 = P_2Q_2 = \ldots = P_nQ_n$ and $N_1 = N_2 = \ldots = N_n$; so that, dropping subscripts, Equation (58) becomes

$$D = \frac{\sqrt{n}RN}{K\sqrt{PQ}}. \tag{65}$$

Any value of g that maximizes D in Equation (65) will also maximize D in (58).

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 D is a maximum when $g_i = 0$ for all i.

When $g_i = 0$, h_i must equal R_ic, and $P_i = 1/2$. 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,* $c = c_o$, *then all items should be of equal difficulty such that half of those examinees whose ability score is* c_o *will answer each item correctly and half will answer it incorrectly.* Strictly speaking, this conclusion has been proved only for the case where all R_i are equal. It appears to be capable of broad generalization, however.

Similar or related conclusions have been reached by a number of writers (33, 32, 12, 1, 8, 34, 4), starting with different premises and proceeding with varying degrees of rigor. Empirical evidence relating to this point is also available (23, 35, 32).

It should be noted that the conclusion reached in the present paper is quite different from the frequent conclusion that, for example, "If 30% of the applicants for a certain kind of work are to be allowed to pass an examination, each item should be sufficiently difficult that approximately 30% of the examinees will know it" (5, 21-22). If a test is to be used to award a single scholarship to the highest-scoring examinee in a group of 100 examinees, the solution would seem to be to use items that will be answered correctly about half the time by the two most able candidates, rather than to use items that will be answered correctly by only one or two per cent of the total group of examinees.

In order to contrast these two different solutions, it may be pointed out that, given that all items are of about 50 per cent difficulty for the two most able candidates, it is still not possible to state what proportion of the total group will answer each item correctly. If, for example, all examinees in the total group are of roughly equal ability, then in this extreme case, almost half of the total group will answer each item correctly. If, on the

other hand, the examinees in the total group are very heterogeneous in ability, then it might be that no one, except for the two most able candidates, would answer any item correctly. It is thus seen to be unreasonable to use the expected proportion of correct answers among the total group of examinees as a basis for determining the optimum item difficulty level; for this proportion varies with the heterogeneity of the group even when the ability of the two most able candidates remains fixed. The optimum item difficulty level must depend on the ability level of the examinees near the cutting score, not on the ability levels of all examinees in the group.

[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 $(1 + 1/k)/2$. Mathematical investigation, which will not be given here, shows that this is clearly not the case, but 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. Similar conclusions have been reached independently by Cronbach and Warrington (4).]

C3. *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 (65). 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 (20). 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:

$$M_{s.c} = nA\left(\frac{M_h - Rc}{G}\right), \tag{66}$$

and

$$\sigma_{s.c}^2 = M_{s.c} - nA_2\left(\frac{M_h - Rc}{G}, \frac{M_h - Rc}{G}; \frac{\sigma_h^2}{G^2}\right). \tag{67}$$

From (66) we find that

$$\frac{\partial}{\partial c} M_{s.c} = \frac{nR}{G} N\left(\frac{M_h - Rc}{G}\right). \tag{68}$$

The discrimination index for a test characterized by given values of R, M_h, and σ_h may be obtained by using (67) and (68) in conjunction with the following formula (*cf.* 57):

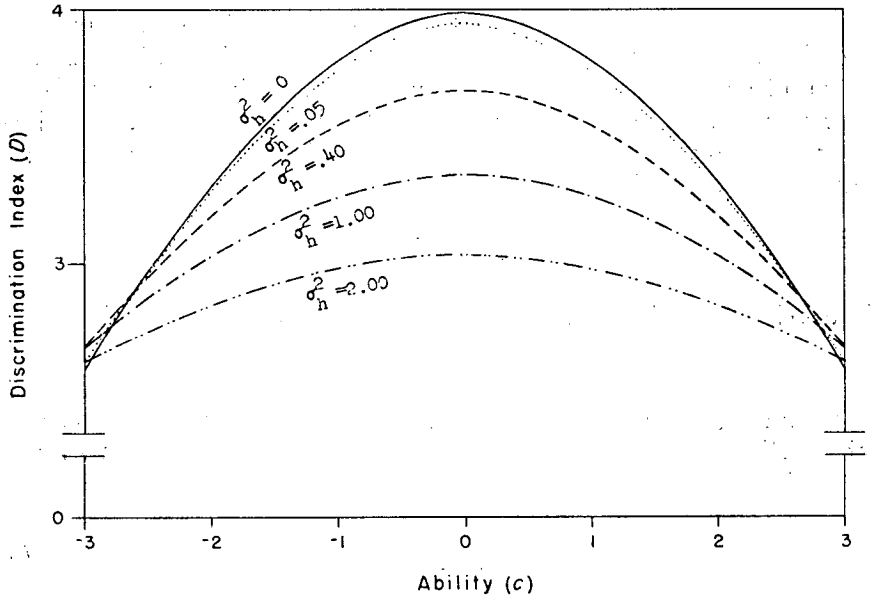
$$D = \frac{1}{\sigma_{s.c}} \frac{\partial}{\partial c} M_{s.c}. \tag{69}$$

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 3 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_i 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—

most ordinary tests, in the author's experience, have values of σ_h^2 between .15 and .50.

Figure 3 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



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

FIGURE 3

our tests could be improved by restricting the range of item difficulty. [The reader may wish to refer to Davis (5) and Flanagan (10) for arguments opposing this point of view.] It is, of course, obvious that the higher the item intercorrelations, the less discriminating the test composed solely of items of 50 per cent difficulty will be for examinees at the extremes of the ability scale. It would seem, however, that in the ordinary type of test the item intercorrelations are not sufficiently high to require a spread of item difficulty values in order to obtain optimum discrimination. Cronbach and Warrington (4) have independently reached similar conclusions with respect to multiple-choice tests.

C4. *Relation of the Discrimination Index to the Test Reliability and to a Certain Maximum Likelihood Statistic*

Let us consider the following function of D and expand it by means of (69), using the symbol b_{sc} for the derivative of $M_{s.c}$ with respect to c:

$$\frac{D^2}{1+D^2} = \frac{b_{sc}^2}{\sigma_{s.c}^2 + b_{sc}^2}. \tag{70}$$

Suppose now that some hypothetical test could be devised such that for this test for some special group the values of b_{sc} and of $\sigma_{s.c}$ could be treated as roughly constant over the range of ability in the group tested. In this hypothetical situation, b_{sc} would be the regression coefficient in the equation for predicting s from c, and consequently, by the usual formula for a regression coefficient,

$$b_{sc} = \frac{\sigma_s}{\sigma_c} r_{sc}. \tag{71}$$

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

$$\sigma_{s.c}^2 = \sigma_s^2(1 - r_{sc}^2). \tag{72}$$

Choosing the units of measurement for c so that $\sigma_c = 1$ and then substituting (71) and (72) in (70), we would obtain for this hypothetical case

$$\frac{D^2}{1+D^2} = r_{sc}^2. \tag{73}$$

If 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,

$$\frac{D^2}{1+D^2} = r_{ss}. \tag{74}$$

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. For example, if a given test has a discriminating power (D) of 3.0 for examinees at a given ability level, it can be stated that the discriminating power of this test for such examinees is the same as the discriminating power that would be achieved at all ability levels by a hypothetical test, with the postulated properties, characterized by a reliability of $r_{ss} = 9/(1 + 9) = .90$.

It may also be noted here that the writer has very recently found for the case of equivalent items that the standard error of the maximum likelihood statistic for estimating an examinee's ability from his responses to the test items (scored 0 or 1) is identically equal to the reciprocal of D. Further work with maximum likelihood methods is in progress.

D. THE FREQUENCY DISTRIBUTION OF TEST SCORES

The univariate distribution of test scores, f_s, may be obtained by integrating (15) according to usual procedures:

$$f_s = \int_{-\infty}^{\infty} f_{cs}dc = \Sigma^* \int_{-\infty}^{\infty} f_c \Pi_s P_i \Pi_{n-s} Q_i dc \quad (s = 0, 1, \ldots, n). \quad (75)$$

Equation (75) is valid for any frequency distribution of ability (f_c) in the group of examinees tested. If c is normally distributed, the equation becomes

$$f_s = \Sigma^* \int_{-\infty}^{\infty} N(c) \Pi_s P_i \Pi_{n-s} Q_i dc \quad (s = 0, 1, \ldots, n). \quad (76)$$

It should be noted that f_s is not a function of c, which here serves merely as a dummy variable.

Unfortunately, the indicated integration cannot in general be performed directly, and consequently no simple algebraic expression for f_s can be obtained. Nor has the writer as yet succeeded in finding any representation of f_s in series form that is sufficiently rapidly convergent to be of much value. Numerical integration of (76) is relatively easy, however, except for the amount of routine work required to calculate the necessary products of P's and Q's. The "theoretical" distributions in Figures 5 through 11 at the end of this monograph serve to illustrate the results obtained when the values of f_s are calculated from (76) for a variety of short tests, using actual observed values of h_i and R_i.

In the special case that occurs when all items are uncorrelated, (76) is seen to reduce to

$$f_s = \Sigma^* \Pi_s p_i \Pi_{n-s} q_i, \quad (77)$$

a generalization of the binomial. If all items are of equal difficulty, we may drop the subscript i and obtain for this special case $f_s = \binom{n}{s} p^s q^{n-s}$, which is the usual binomial.

Another special case that is of interest occurs when all items are of 50 per cent difficulty ($h = 0$) and all values of R_i are equal to $1/\sqrt{2}$. In this case $g_i = -c$, and (76) becomes

$$f_s = \binom{n}{s} \int_{-\infty}^{\infty} N(c) A^s(-c) B^{n-s}(-c)\, dc \quad (s = 0, 1, \ldots, n), \quad (78)$$

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

$$\frac{d}{dc} A(-c) = N(c); \quad (79)$$

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 (78) by dA, we obtain the result

$$f_s = \binom{n}{s} \int_0^1 A^s (1 - A)^{n-s} dA. \quad (80)$$

The integral in (80) is a beta function; so finally,

$$f_s = \frac{n!}{s!(n-s)!} \frac{s!(n-s)!}{(n+1)!} = \frac{1}{n+1} \quad (s = 0, 1, \ldots, n). \quad (81)$$

Since s can vary only 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 (8). In a previous article (15), 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 (76) will give such U-shaped distributions whenever the item intercorrelations are sufficiently high. Apparently, when c is normally distributed, R_i must be greater than $1/\sqrt{2}$ before this can occur, however. Such large values of R_i 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 (76), 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:

s	f_s	cumulative frequency
5	.25	1.00
4	.00	.75
3	.25	.75
2	.32	.50
1	.16	.18
0	.02	.02
	1.00	

Referring to a table of the normal curve, the required cumulative frequencies are found to correspond to relative deviates of ∞, $+0.67$, $+0.67$, $+0.00$, -0.92, and -2.05. A little thought will show that if five items are selected that are perfectly correlated with 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 (76). 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 (76), 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 (77), 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 n times the mean item difficulty:

$$M_s = \Sigma p_i. \tag{82}$$

This same result may be derived from (76), as follows:

$$M_s = \sum_{s=0}^{n} s f_s = \int_{-\infty}^{\infty} N(c) \Sigma_s s \Sigma^* \Pi_s P_i \Pi_{n-s} Q_i dc = \int_{-\infty}^{\infty} N(c) M_{s.c} dc. \tag{83}$$

Substituting the value for $M_{s.c}$ from (16), we have

$$M_s = \sum_i \int_{-\infty}^{\infty} N(c) P_i dc. \tag{84}$$

This integral may be evaluated with a result that agrees with that obtained algebraically and presented in (82).

The standard deviation of the test score distribution may likewise be derived without assumptions by simple algebra, as shown in (30). This same result also may be obtained from (76) by evaluation of the integrals in the formula

$$\sigma_s^2 = \sum_{s=0}^{n} s^2 f_s - M_s^2. \tag{85}$$

The third and higher-order moments of the test score distribution cannot be expressed as simple functions of the item difficulties and intercorrelations. These moments can be obtained by expanding the necessary integrals by means of the infinite series developed by Pearson (28), but the result is too cumbersome to be of much practical value for present purposes. For any given numerical case, these moments can be obtained by numerical integration, as has been done for certain tests in Part III of the present monograph.

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 a formula for the frequency distribution of test scores has been obtained, to note that the line of reasoning followed for this purpose is, in broad outline, the same as that used by Lazarsfeld (22) 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

E1. *Derivation of the Distribution*

In discussing the limiting distribution of scores as n becomes very large, we shall 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.

It was seen from (25) 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,

$$t = M_{z.c}. \tag{86}$$

In the case of relative scores, the regression is seen from (16) to be given by the average value of P:

$$M_{z.c} = \frac{1}{n}\Sigma P_i = M_P. \tag{87}$$

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. It is seen from (86) and (87) that the distribution of t is the same as the distribution of the function $M(c)$. If it is assumed that $R_i \neq 0$, the inverse function of $M(c)$ may be denoted 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

$$f_t = f[M^{-1}(t)]\frac{\partial M^{-1}(t)}{\partial t}. \tag{88}$$

When c is normally distributed, we have

$$f_t = N[M^{-1}(t)]\frac{\partial M^{-1}(t)}{\partial t}. \tag{89}$$

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

$$t = M(c) = \frac{1}{m}\sum_{i=1}^{m} P_i. \tag{90}$$

We find that (*cf.* Equation 17)

$$\frac{\partial t}{\partial c} = \frac{1}{m}\sum_{i=1}^{m}\frac{R_i N(g_i)}{K_i}. \tag{91}$$

Since $c = M^{-1}(t)$, $\dfrac{\partial M^{-1}(t)}{\partial t}$ is the reciprocal of the expression on the right side of (91). Substituting this result in (89), we obtain finally:

$$f_t = \frac{mN[M^{-1}(t)]}{\sum^{m} \frac{R_i N(g_i)}{K_i}}. \tag{92}$$

In (92), g_i is as usual (9) a function of h_i, R_i, and c; but here c is treated as a function of t, so that g_i is a function of t. For purposes of calculation, f_t may perhaps best be computed from the parametric equations:

$$f_t = \frac{mN(c)}{\sum^{m} \frac{R_i N(g_i)}{K_i}}, \quad t = M(c), \tag{93}$$

where c is the parameter. When the values of h_i and R_i 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 shall denote by $F(t)$, may be of interest; it is found from (89), by the usual methods, to be

$$F(t) = \int_{-\infty}^{t} f_t dt = \int_{-\infty}^{M^{-1}(t)} N(c)\,dc. \tag{94}$$

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 (94). Equation (94) is of particular interest, since it shows clearly that the distribution of t can never be strictly normal. The distribution will be approximately normal, however, whenever $M(c)$ is an approximately linear function of c over the range in which most of the cases occur. It is thus seen that t will be approximately normally distributed in just those cases where the regression of s on c is practically linear over the range of the actual data, as discussed in detail in Section B8.

The frequency distribution of relative scores actually approaches the frequency distribution f_t as a limit when n becomes infinitely large. This same result may be obtained from (76) by finding the limit of the characteristic function as 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 n, and f_t.

We may note at this point that, because of the relation between t and 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 (12).

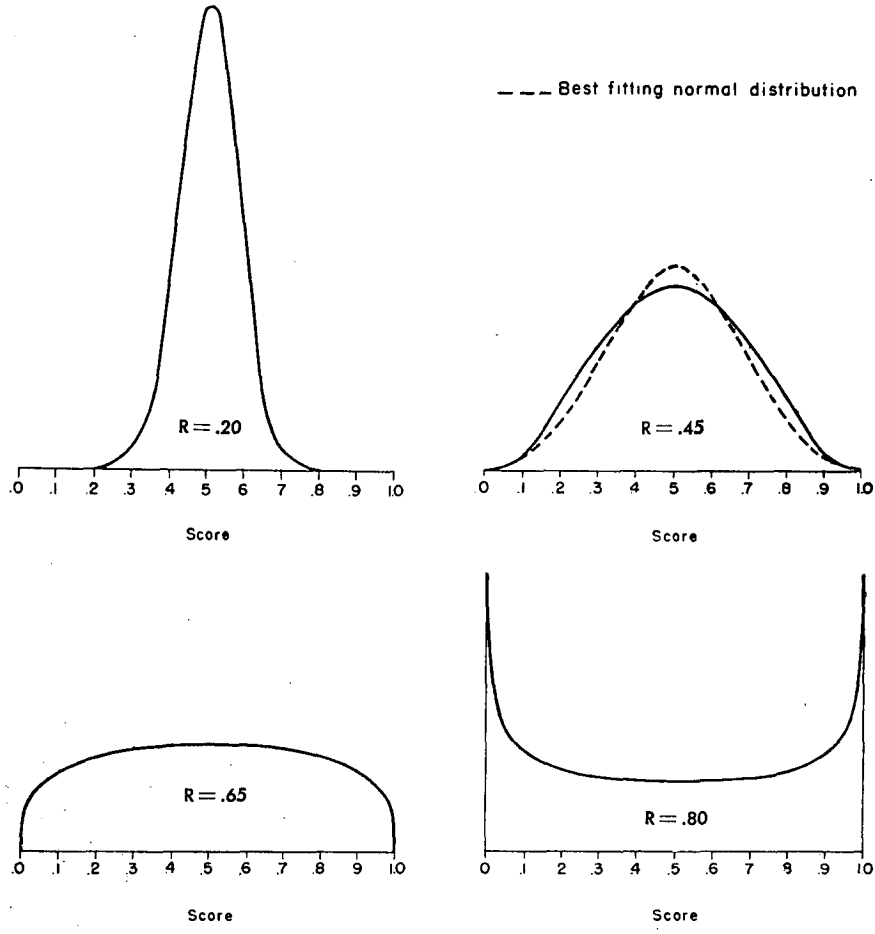
E2. *Illustrative Examples of the Distribution of Relative True Scores*

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 4 illustrates the effect on the shape of the true-score distribution of increasing the correlation (R) of the items with ability. The values of f_t needed for drawing these distributions were obtained from Equation (92). 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 roughly 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 (81). 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, if sufficiently long, 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.



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

FIGURE 4

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 independent when c is fixed. The conditional bivariate distribution of test scores will therefore be

$$f_{s_1 s_2 . c} = f_{s_1 . c} f_{s_2 . c} . \tag{95}$$

The right side of this equation may be evaluated by (12).

The trivariate distribution of s_1, s_2, and c is therefore

$$f_{cs_1 s_2} = N(c) f_{s_1 s_2 . c}; \tag{96}$$

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

$$f_{s_1 s_2} = \int_{-\infty}^{\infty} N(c) f_{s_1 s_2 . c} dc . \tag{97}$$

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, (97) 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 11 to 17 at the end of this monograph.

In the special case where all items are of 50 per cent difficulty and $R_i = \sqrt{.5}$ for all items, the bivariate distribution of test scores is found by direct integration (*cf.* 81) to be

$$f_{s_1 s_2} = \frac{1}{n_1 + n_2 + 1} \frac{\binom{n_1}{s_1}\binom{n_2}{s_2}}{\binom{n_1 + n_2}{s_1 + s_2}} . \tag{98}$$

[It may be noted that the second fraction in (98) is the general term of a hypergeometric distribution.] The bivariate distribution of (98) is somewhat of a curiosity. The marginal distributions of s_1 and s_2 are of course rectangular (81). The distribution of $s_1 + s_2$, obtained by summing the frequencies in (98) 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 (98) may be well characterized as saddle-shaped.

The regression of s_2 on s_1 for the special frequency surface of (98) is found to be

$$M_{s_2 . s_1} = \frac{n_2 (s_1 + 1)}{n_1 + 2} . \tag{99}$$

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 seems reasonable for such a case as the Sonar Pitch Memory Test with which Cronbach and Warrington are concerned. In the case of most ordinary multiple-choice tests, however, this assumption is certainly only a rough approximation to the real life situation and should not be expected to yield as good agreement with empirical results as would be found in work with free-response items.

PART 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, 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 (76) and (97), 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 (76) and (97); 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 insofar 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, was at least approximately normally distributed in the group tested. The success of the empirical verification undertaken may depend to a considerable extent upon 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 (31). 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 three 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; or (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

J1. *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 at the end of this monograph. 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. (The allocation of the 28 items to various tests, as indicated by Table 5, will be discussed in the next section.)

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 with the sampling error of a tetrachoric correlation when the sample size is 136 and the true correlation is .30. Table 1

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

Item Difficulty	Item Difficulty .16	.50	.84
.16	.20	.15	.17
.50	.15	.13	.15
.84	.17	.15	.20

gives a few values of this sampling error for various combinations of item difficulties, as computed from the table provided by Hayes (14).

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 estimated diagonal entries had been inserted in the residual matrix. The sum of squares of the unrotated second factor loadings was found to be only 1.6 as compared with 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 derivations in the preceding sections. This difficulty presumably could have been avoided if more cases had been available from which to compute the tetrachoric correlations.

J2. *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 5, together with the difficulties and common factor loadings of the items. The eight tests may be briefly summarized as follows:

Test 2 — difficult items ($p_i = .09$ to $p_i = .34$).
Test 5 — items of medium difficulty ($p_i = .43$ to $p_i = .68$).
Test 8 — easy items ($p_i = .71$ to $p_i = .90$).
Test 8′ — like Test 8, but slightly shorter.
Test 82 — five "easy" and five "difficult" items.
Test h — highly discriminating items, irrespective of difficulty ($R_i = .55$ or more).
Test L — poorly discriminating items, irrespective of difficulty ($R_i = .47$ or less).
Test r — item difficulties rectangularly distributed.

The univariate frequency distribution of scores was investigated for each test. 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.

J3. *Obtaining the Theoretical Bivariate Frequency Distribution of Test Score and Ability*

Fifteen values of 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 c from (7), (8), (9), and (10).

For each test, for each value of c, the theoretical conditional distribution of test scores was calculated from (12). An example for $n = 3$ will clarify the method used. Suppose (for a given value of 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

$$(.5 + .5)(.4 + .6)(.2 + .8) = (.2 + .5 + .3)(.2 + .8) = .04 + .26 + .46 + .24.$$

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 s for each of the 15 values of 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 s and c is seen from (14) to be

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

Tables 7, 8, and 9 present, for the selected values of c, the theoretical ordinates of the bivariate distribution of test score and ability for three of the eight tests, selected for illustrative purposes. It should be noted that the values given in the table are ordinates for the stated values of c, not cell frequencies corresponding to class intervals of c; consequently the sum of a row of tabled values does not give the marginal frequency of the test score.

J4. *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 (76). It was found most satisfactory in the present case to perform the numerical integration by use of the trapezoidal rule:

$$\int_{c_0}^{c_m} f dc = h\left(\frac{f_0}{2} + f_1 + f_2 + \cdots + \frac{f_m}{2}\right)$$

approximately, where f is the function of c to be integrated, f_u is the value of f when $c = c_u$, c_m is the upper limit of integration, c_0 is the lower limit of integration, and h is the difference between any two successive values of c. In the present case, c_0 and c_m may be taken as -4.0 and $+4.0$, respectively, so that the ordinates f_0 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

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

approximately. The adequacy of this method of numerical integration for present purposes was carefully checked, especially by application of Simpson's one-third rule.

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 for three tests in the right-hand columns of Tables 7-9. Tables 11-17 show the frequency distributions for all eight tests after multiplication of the relative frequencies by the number of actual examinees, 136. The comparison of these theoretical distributions with the corresponding actual distributions will be discussed at a later point.

J5. *Obtaining the Theoretical Bivariate Frequency Distribution for Two Subtest Scores*

The theoretical bivariate distribution for two subtest scores may be obtained from (97). 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 (97) is carried out by means of the trapezoidal rule, we have

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

where the primed matrices are transposed.

The theoretical distributions of test scores obtained in this way are presented in Tables 11-17 and discussed in Section K2.

K. COMPARISON OF THEORETICAL AND ACTUAL RESULTS

K1. *Comparison of Univariate Frequency Distributions of Test Scores*

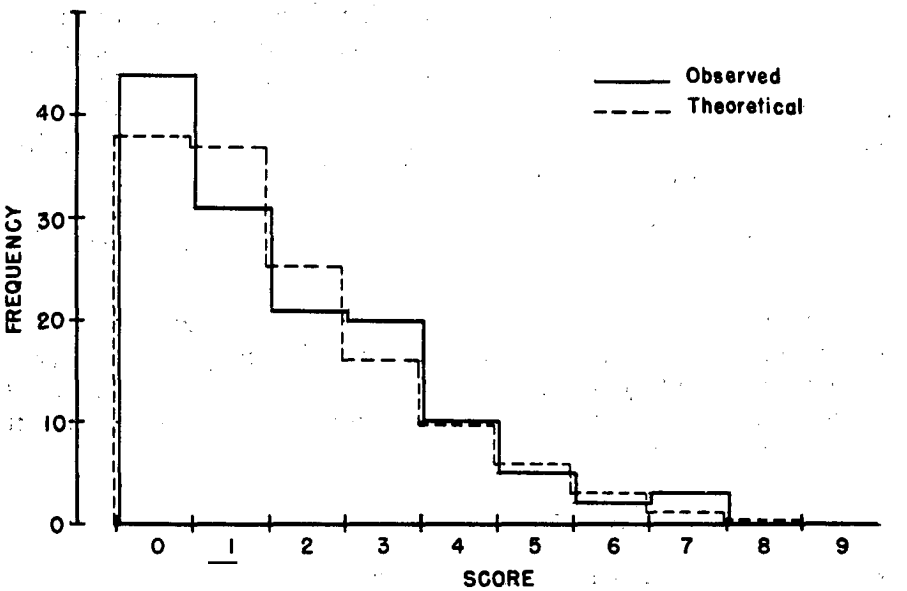
The theoretical and obtained univariate distributions of scores (except for Test 8′, which is very similar to Test 8) are compared graphically in Figures 5 through 11, and numerically in the margins of Tables 11 through 17. Table 10 shows the means and standard deviations of all the distributions, together with a measure of skewness,

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

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

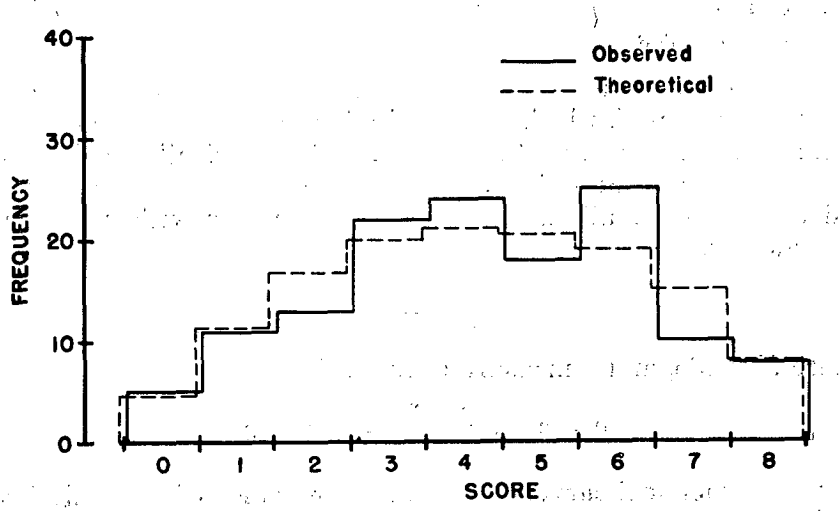
$$\beta_2 - 3 = \frac{1}{N\sigma_s^4} \sum^N (s - M_s)^4 - 3.$$

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.



Theoretical and Observed Distributions of Scores on Test 2 (Difficult Items)

FIGURE 5



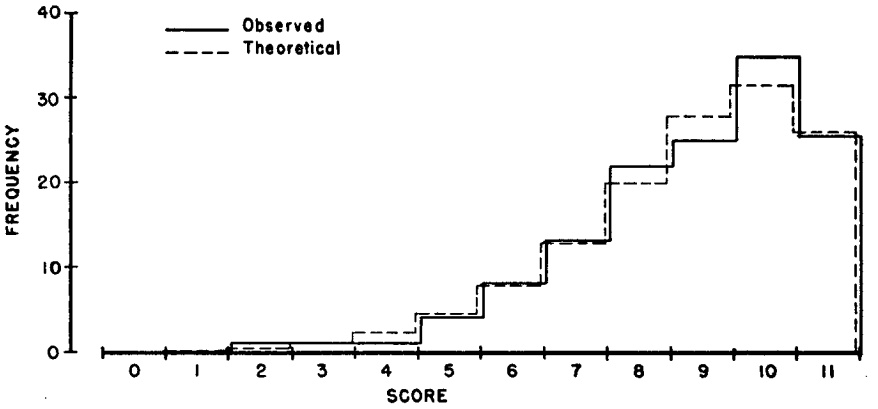
Theoretical and Observed Distributions of Scores on Test 5
(Items of Medium Difficulty)

FIGURE 6

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 about 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 items of medium difficulty, 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.

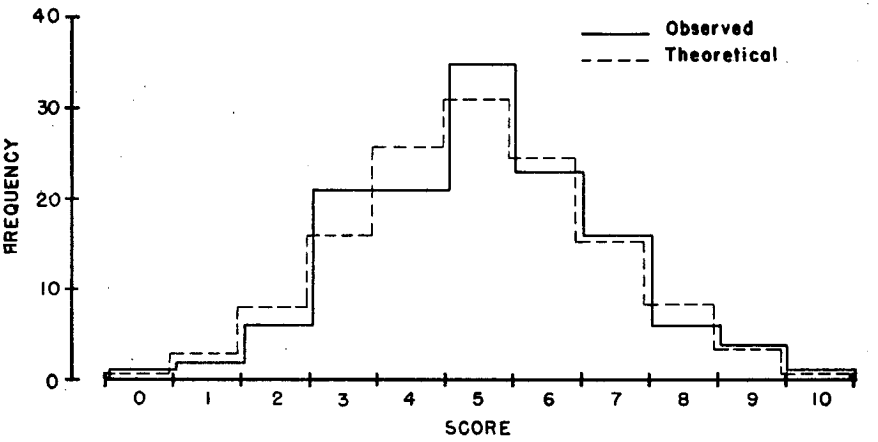
The standard error of α_3 for a sample of 136 cases drawn at random from a normal population is 0.21, the standard error of $\beta_2 - 3$ in the same situation 0.41. Several of the distributions of test scores presented here are distinctly nonnormal, however, and these standard errors cannot be considered 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 may 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 statistically significant differences are bound to be found if only enough cases are used. The real issue is whether



Theoretical and Observed Distributions of Scores on Test 8 (Easy Items)

FIGURE 7



Theoretical and Observed Distributions of Scores on Test 82 (Five Easy and Five Difficult Items)

FIGURE 8

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 it is borne in mind that no significance test is being made. 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 Table 10. In the calculation of 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 probabilities that the obtained values of chi-square would be exceeded in random sampling are given in Table 2

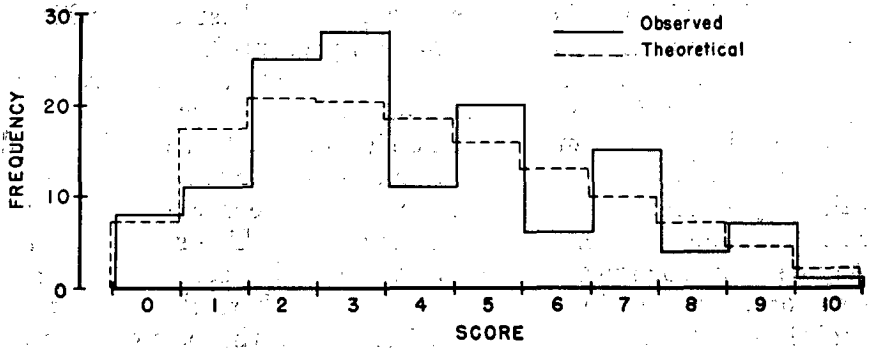
TABLE 2

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

Chi-Square	*Degrees of Freedom*	*Probability*
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

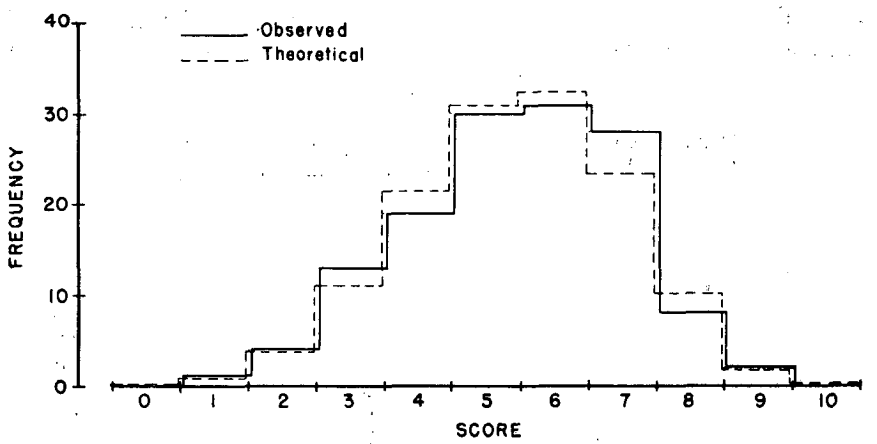
for the case where the population parameters are not estimated from the data. These probabilities are higher than would be appropriate if the difference between theoretical and observed frequency distributions were being tested for significance.

The only test that has a rather large value of chi-square is Test *h*. The theoretical and observed distributions of scores on



Theoretical and Observed Distributions of Scores of Test h
(Highly Discriminating Items)

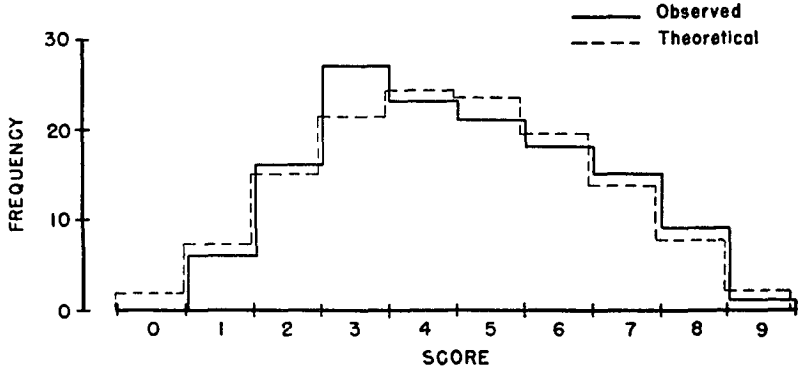
FIGURE 9



Theoretical and Observed Distributions of Scores on Test L
(Poorly Discriminating Items)

FIGURE 10

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. The conclusion that these irregularities are the result of sampling fluctuations and are not due to peculiarities of the test items is borne out by the fact that another group of examinees, drawn at random from the same population, obtained a distribution of scores on the same test that shows no trace of the toothed appearance of the actual distribution in Figure 9.



Theoretical and Observed Distributions of Scores on Test *r*
(Rectangular Distribution of Item Difficulty)

FIGURE 11

K2. *Comparison of Bivariate Frequency Distributions of Test Scores*

The theoretical and observed bivariate distributions of scores on the pairs of tests investigated are compared in Tables 11 through 17. The dotted lines in the tables indicate how grouping was carried out for the computation of chi-square. 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.

As before, no significance test is possible with these data. Table 3, however, gives probabilities for the obtained values of chi-square

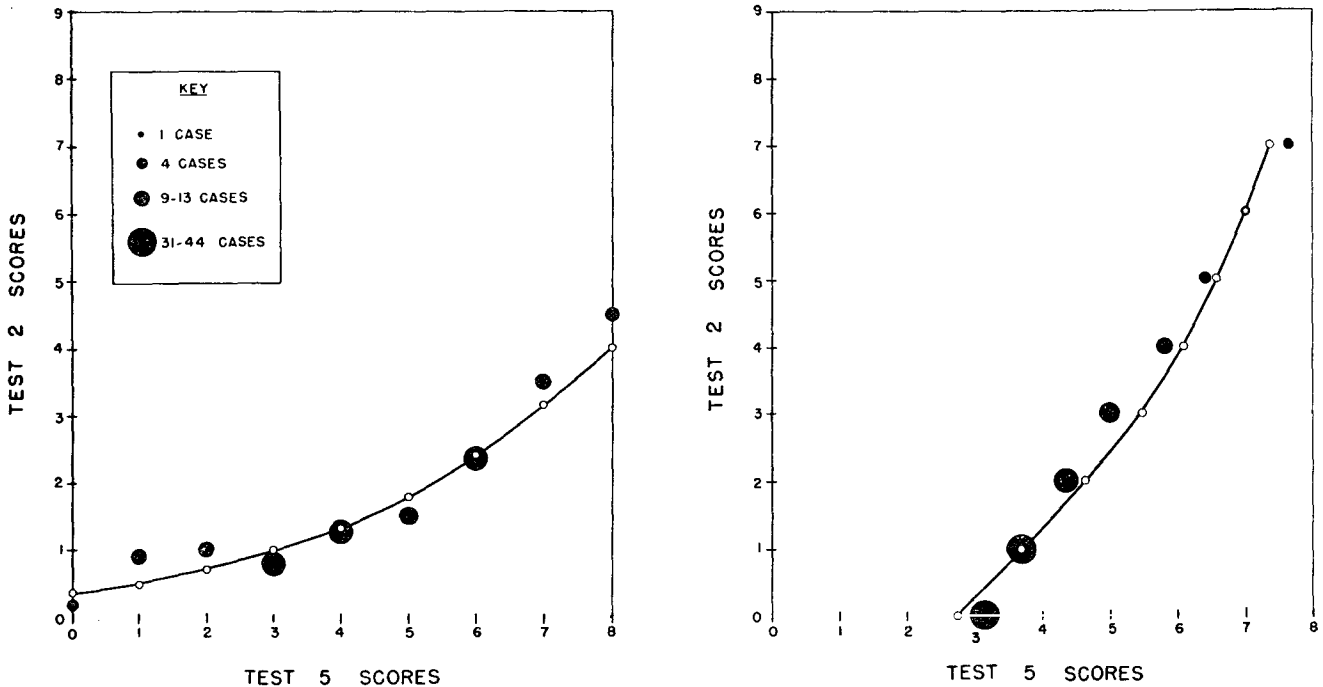
TABLE 3

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

Chi-Square	*Degrees of Freedom*	*Probability*
.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

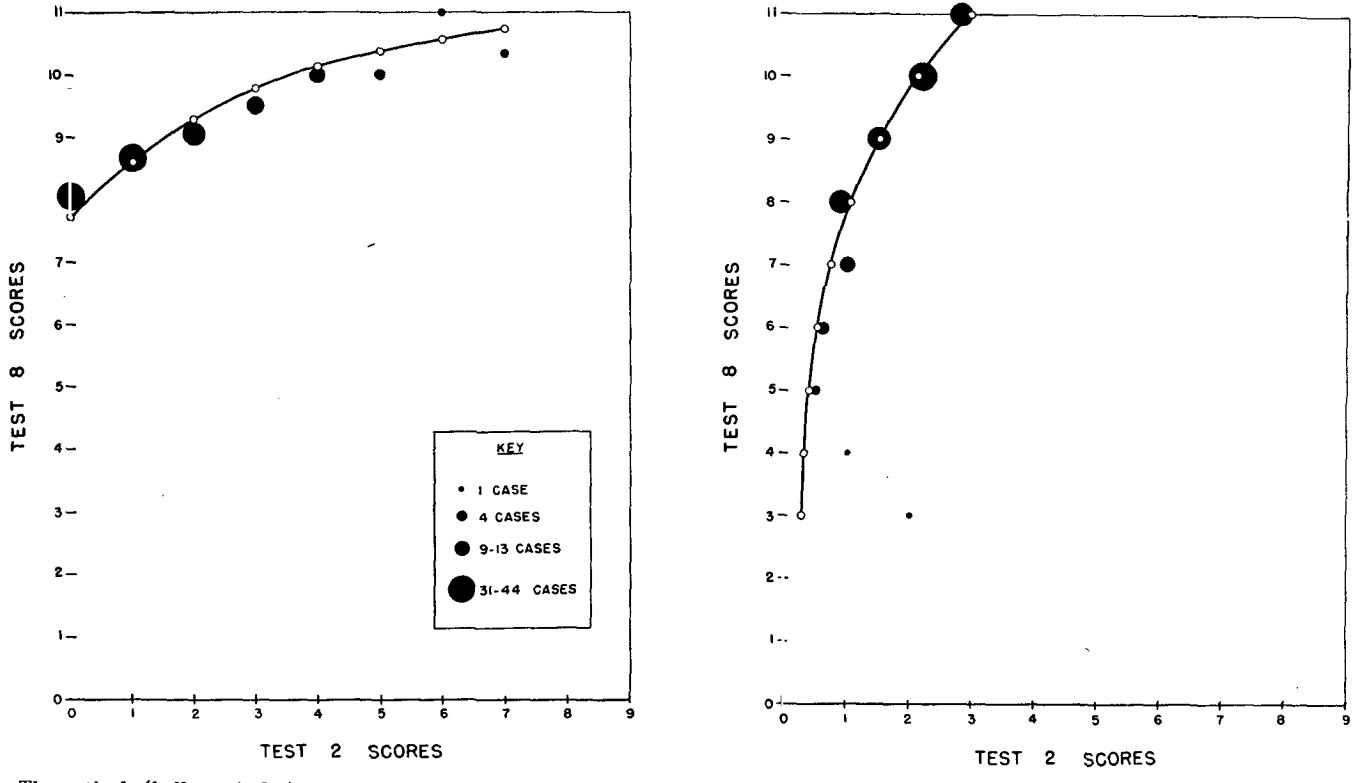
that may be interpreted in the same fashion as the probabilities given in Table 2. It is seen that each of the theoretical distributions provides an adequate fit to the corresponding observed distribution.

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 11 through 17. 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 described in Section J4. No further statistics relating to these regression lines have been calculated. Judging from the diagrams, however, we may say that the theoretical regressions seem to provide a good fit to the observed data.



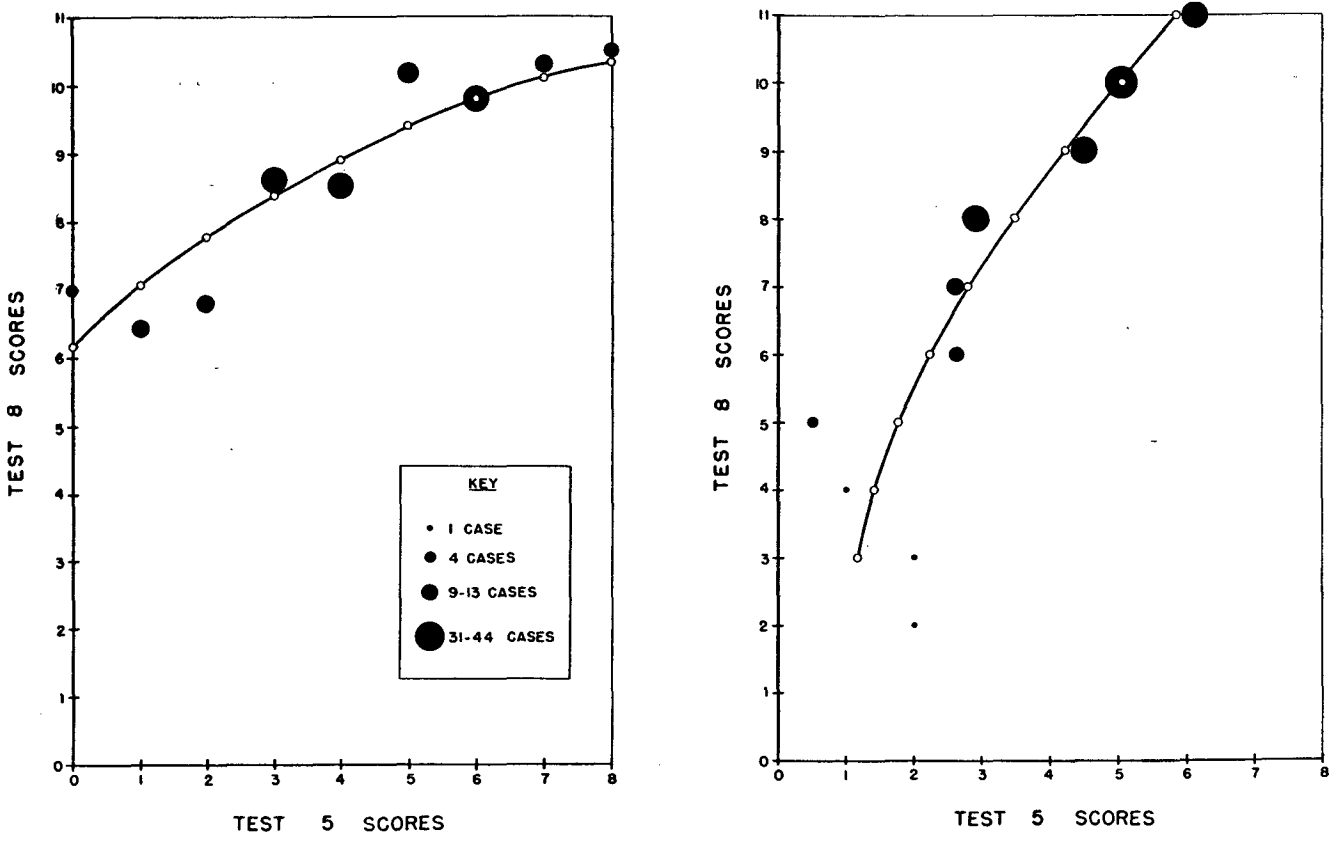
Theoretical (hollow circles) and Actual (solid circles) Regressions for Test 2 on Test 5 (left) and for Test 5 on Test 2 (right)

FIGURE 12



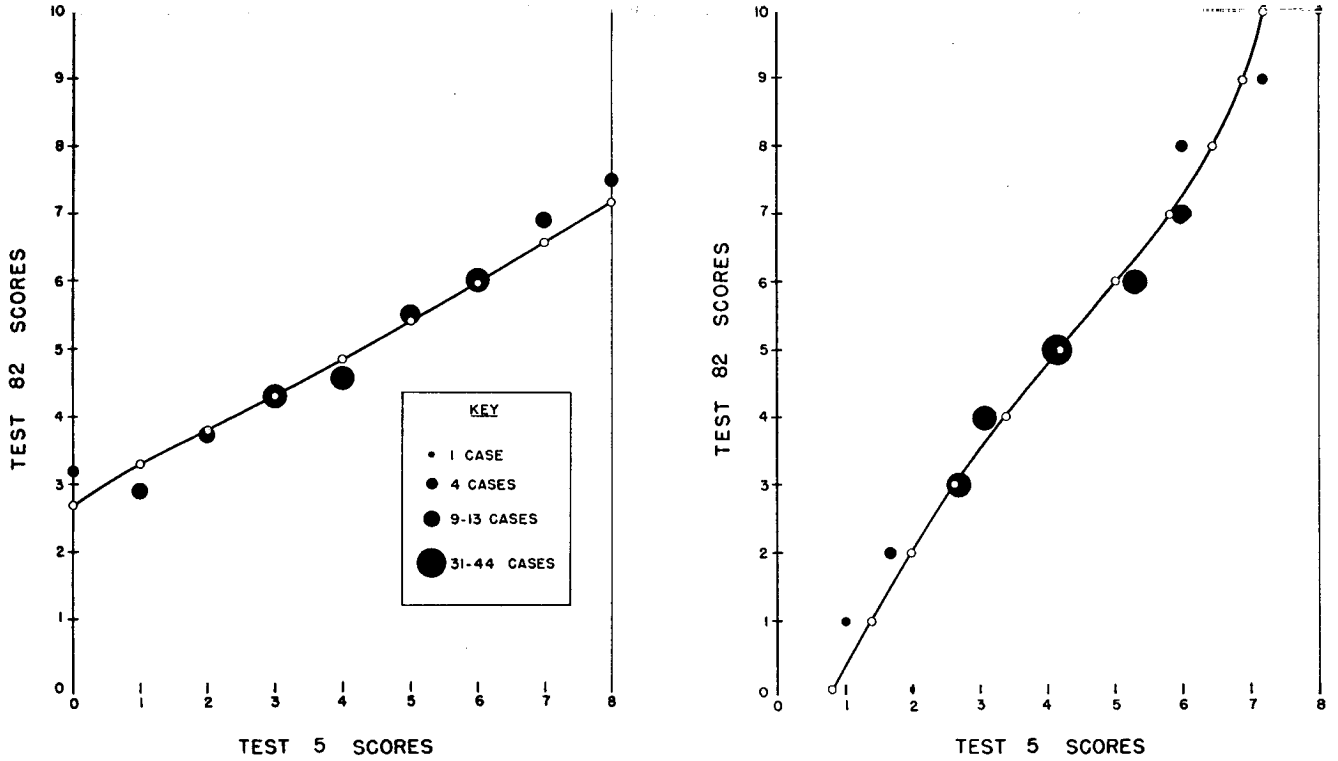
Theoretical (hollow circles) and Actual (solid circles) Regressions for Test 8 on Test 2 (left) and for Test 2 on Test 8 (right)

FIGURE 13



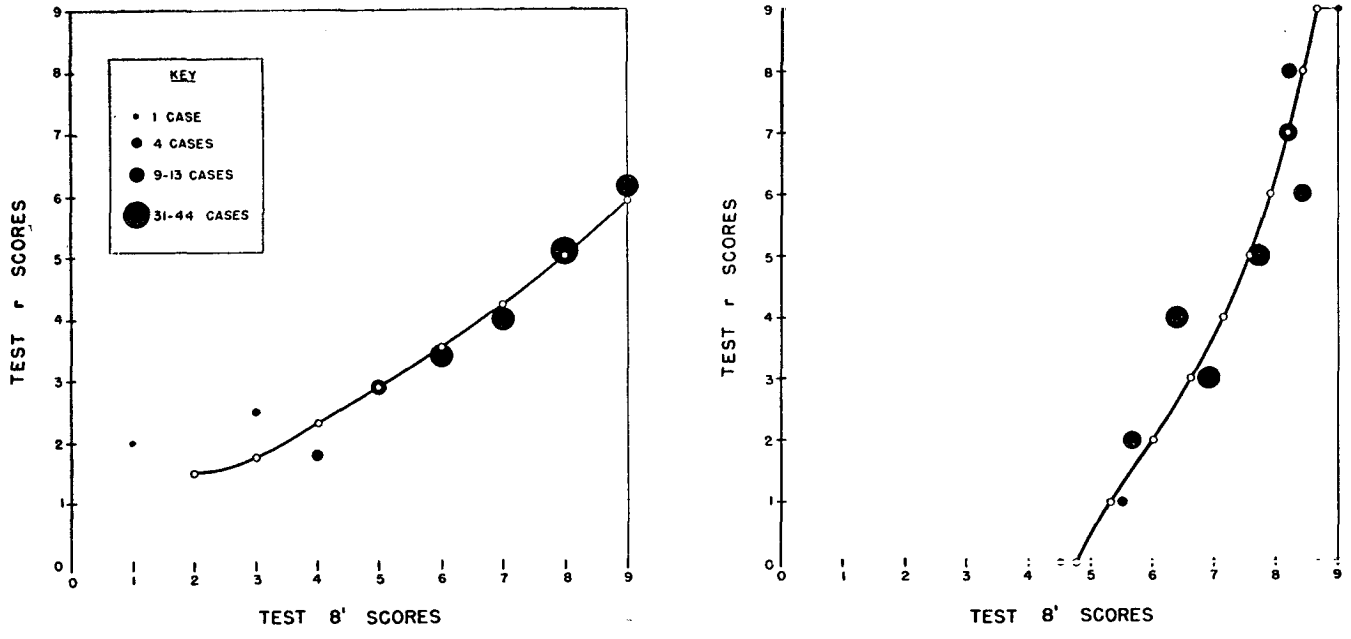
Theoretical (hollow circles) and Actual (solid circles) Regressions for Test 8 on Test 5 (left) and for Test 5 on Test 8 (right)

FIGURE 14



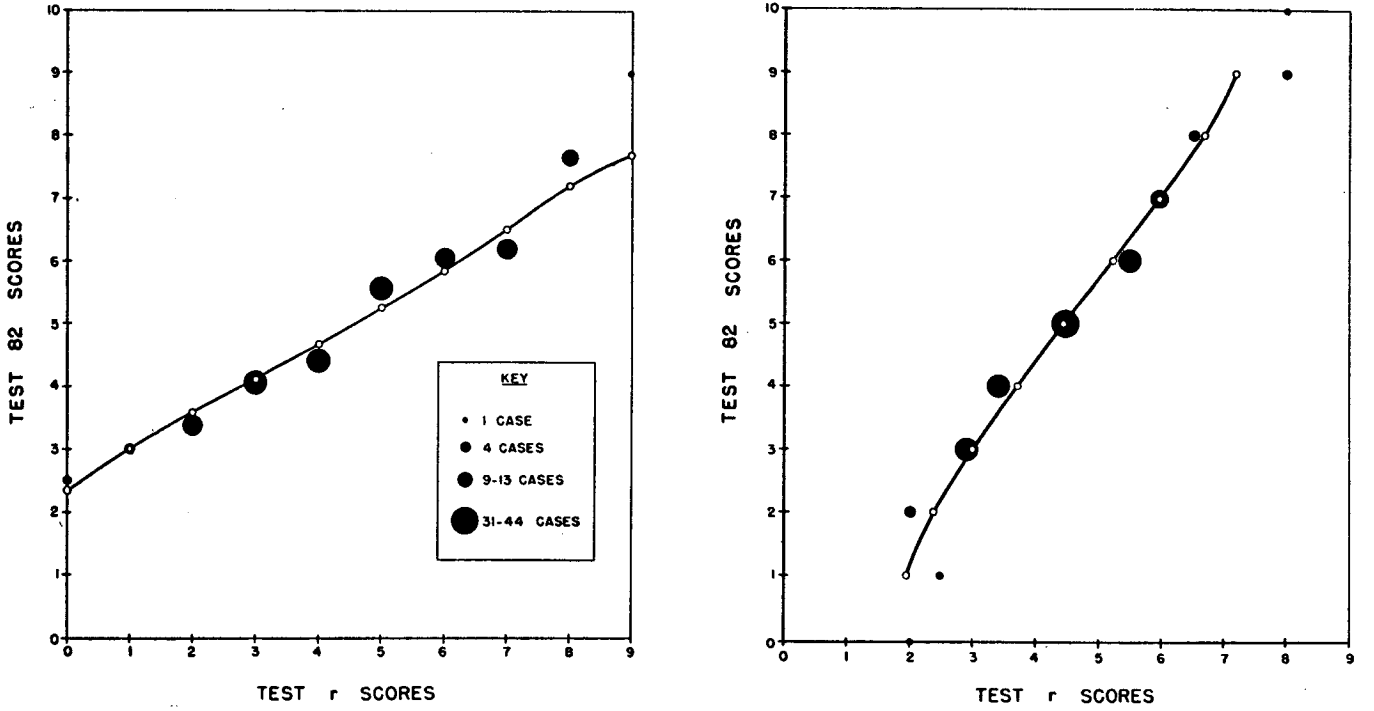
Theoretical (hollow circles) and Actual (solid circles) Regressions for Test 82 on Test 5 (left) and for Test 5 on Test 82 (right)

FIGURE 15



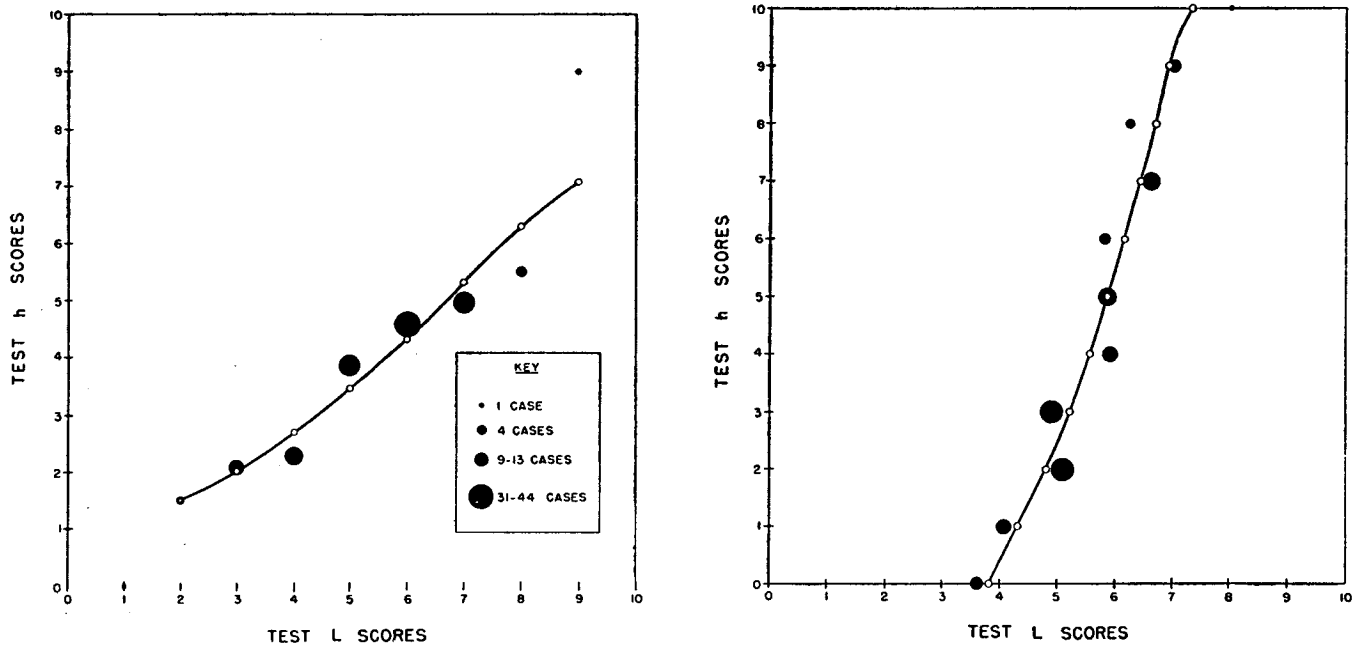
Theoretical (hollow circles) and Actual (solid circles) Regressions for Test r on Test 8′ (left) and for Test 8′ on Test r (right)

FIGURE 16



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

FIGURE 17



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

FIGURE 18

PART IV

SUMMARY AND CONCLUSIONS

The theory developed here 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 aids in the prediction of 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 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 frequen-

cies 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 have been 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 derived, 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), and for the curvilinear correlation of test score on ability (Section B7). 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 leads 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 vary 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 free-response test composed entirely of items of fifty per cent difficulty will, in terms of the discrimination index developed here (58), 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 3).

(6) The shape of the frequency distribution of test scores or of true scores (Sections D and E, and also Figures 4-11) 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 distributions has been urged more than once in the recent literature (**8, 15**); 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 been derived here only for the case where the test items cannot be answered correctly by guessing. The case where guessing occurs has been formulated in detail but is not presented here.

An empirical study of seven short tests has shown very good agreement between the univariate and bivariate test score distributions actually obtained and the corresponding theoretical distributions predicted from the item difficulties and intercorrelations by means of the formulas developed for this purpose.

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	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′	33		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′	50	28		44	15	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	44		32	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	*	*	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	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	*	*	15	20		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	04	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	19	17	00	−07	−01	23	00	24	32	16	36	−05	32	08	33	13	30
10	62	34	38	30	20	22	59	42	04	15		34	38	30	30	34	01	52	28	27	24	47	23	44	−11	45	35	21
11	20	52	25	47	30	25	53	58	19	19	34		17	09	28	−09	10	28	24	24	39	42	19	51	−17	33	38	37
13	29	12	56	25	13	−03	21	07	17	17	38	17		23	08	08	18	37	33	12	18	48	33	52	04	14	33	21
16	10	22	22	41	41	00	33	25	−06	00	30	09	23		02	09	03	15	35	22	30	10	27	13	−02	20	23	09
18	19	28	22	14	14	*	33	−08	−07	−07	30	28	08	02		07	10	25	10	39	18	26	30	13	−06	41	12	23
20	22	−10	40	19	−02	19	36	22	07	−01	34	−09	08	09	07		16	21	24	07	08	22	43	39	01	17	28	02
39	02	09	−05	−16	24	−18	10	10	04	23	01	10	18	03	10	16		11	00	−08	−01	30	17	23	08	−12	−29	14
40	56	05	32	51	30	21	28	25	11	00	52	28	37	15	25	21	11		33	40	20	40	32	45	21	33	39	19
44	26	18	51	43	24	25	50	24	17	24	28	24	33	35	10	24	00	33		−05	32	26	38	39	−31	43	21	41
45	41	43	42	31	21	24	06	38	01	32	27	24	12	22	39	07	−08	40	−05		26	56	33	35	−28	26	33	17
46	24	14	10	27	27	17	19	29	28	16	24	39	18	30	18	08	−01	20	32	26		03	12	40	−11	47	08	50
48	36	54	34	12	34	41	45	48	08	36	47	42	48	10	26	22	30	40	26	56	03		49	58	15	48	26	25
50	27	24	04	32	51	29	36	10	15	−05	23	19	33	27	30	43	17	32	38	33	12	49		39	−23	46	54	−03
53	69	38	36	41	41	48	51	36	26	32	44	51	52	13	13	39	23	45	39	35	40	58	39		−04	53	26	38
54	−08	34	−06	−30	−13	−02	−05	*	05	08	−11	−17	04	−02	−06	01	08	21	−31	−28	−11	15	−23	−04		00	−15	−07
55	*	28	12	31	*	48	42	14	30	33	45	33	14	20	41	17	−12	33	43	26	47	48	46	53	00		38	42
56	08	16	33	25	43	10	56	12	17	13	35	38	33	23	12	28	−29	39	21	33	08	26	54	26	−15	38		29
57	37	10	26	04	35	10	20	41	19	30	21	37	21	09	23	02	14	19	41	17	50	25	−03	38	−07	42	29	

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

TABLE 5

Item Difficulties and Common Factor Loadings for the Items Included in Each of the Eight Tests

Item No.	Item Difficulty	Common Factor Loadings of Test Items							
		Test 2	Test 5	Test 8	Test 8′	Test 82	Test *h*	Test *L*	Test *r*
54	.088	—.093	...	...	...	...	...	—.093	...
6	.096	.490	...	...	...	...	...	...	.490
18	.096	.332	...	...	...	.332	...	.332	...
55	.162	.655	...	...	...	.655	.655	...	...
48	.176	.715	...	...	...	.715	.715	...	...
5	.199	.717	...	...	...	...	.717	...	.717
9	.272	.409	...	...	...	.409	...	.409	...
11	.272	.581	...	...	...	.581	.581	...	...
3′	.338	.549	...	...	...	...	.549	...	.549
40	.434	...	.593	...	...	...	.593	...	.593
53	.441	...	.800	...	...	...	.800	...	...
5′	.471	...	.595	...	...	...	.595	...	.595
46	.485	...	.437	...	...	...	...	.437	...
16	.522	...	.352	...	...	...	...	.352	...
10	.574	...	.640	...	...	...	.640	...	.640
56	.610	...	.483	...	...	...	...	...	...
3	.676	...	.476	...	...	...	...	...	.476
44	.713	...	...	.531	.531	.531	...	...	...
57	.721	...	...	.465	.465	...	...	.465	...
20	.735	...	...	.324	.324	...	...	.324	...
45	.750	...	...	.485	.485	.485	...	...	...
50	.801	...	...	.530	...	...	...	...	.530
8	.801	...	...	.299	.299	...	...	.299	...
13	.809	...	...	.466	.466	.466	...	.466	...
1′	.868	...	...	.654	.654	.654	.654	...	...
1	.882	...	...	.481	.481	.481	...	...	...
2	.882	...	...	.495	...	...	...	...	.495
39	.897	...	...	.114	.114	...	...	.114	...
No. of items:		9	8	11	9	10	10	10	9

TABLE 6

Residual Correlations Among Twenty-Eight 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′		−03	11	−06	−06	−02	−07	*	16	23	20	−18	−02	−13	−03	01	−06	17	−09	09	−05	−11	−08	17	−02	*	−24	07
3′	−03		−05	−11	−12	26	07	01	−00	28	−01	20	−14	03	10	−28	03	−28	−11	16	−10	15	−05	−06	39	−08	−10	−16
5′	11	−05		15	−14	−04	10	01	00	14	−00	−10	28	01	02	21	−12	−03	19	13	−16	−08	−28	−12	−00	−27	04	−02
1	−06	−11	15		08	04	*	*	−16	−17	−01	19	03	24	−02	03	−22	22	18	08	06	−22	06	02	−26	−00	02	−18
2	−06	−12	−14	08		−04	*	*	07	−28	−12	01	−10	24	−02	−18	18	01	−02	−03	05	−01	25	01	−08	*	19	12
3	−02	26	−04	04	−04		03	−08	22	32	−08	−03	−25	−17	*	04	−23	−07	−00	01	−04	07	04	10	02	17	−13	−12
5	−07	07	10	*	*	03		−15	15	−02	13	11	−12	08	09	13	02	−14	12	−29	−12	−06	−02	−06	02	−05	21	−13
6	*	01	01	*	*	−08	−15		−05	10	11	30	−16	08	−24	06	04	−04	−02	14	08	13	−16	−03	*	−18	−12	18
8	16	−00	00	−16	07	22	15	−05		−02	−15	02	03	−16	−17	−03	01	−07	01	−14	15	−13	−01	02	08	10	03	05
9	23	28	14	−17	−28	32	−02	10	−02		−11	−05	−02	−14	−21	−14	18	−24	02	12	−02	07	−27	−01	12	06	−07	11
10	20	−01	−00	−01	−12	−08	13	11	−15	−11		−03	08	08	09	13	−06	14	−06	−04	−04	01	−11	−07	−05	03	04	−09
11	−18	20	−10	19	01	−03	11	30	02	−05	−03		−10	−12	09	−28	03	−06	−07	−04	14	00	−12	04	−12	−05	10	10
13	−02	−14	28	03	−10	−25	−12	−16	03	−02	08	−10		07	−08	−07	13	09	08	−11	−02	15	08	15	08	−16	10	−01
16	−13	03	01	24	24	−17	08	08	−16	−14	08	−12	07		−10	−02	−01	−06	16	05	15	−15	08	−15	01	−03	06	−07
18	−03	10	02	−02	−02	*	09	−24	−17	−21	09	09	−08	−10		−04	06	05	−08	23	04	02	12	−14	−03	19	−04	08
20	01	−28	21	03	−18	04	13	06	−03	−14	13	−28	−07	−02	−04		12	02	07	−09	−06	−01	26	13	04	−04	12	−13
39	−06	03	−12	−22	18	−23	02	04	01	18	−06	03	13	−01	06	12		04	−06	−14	−06	22	11	14	09	−20	−34	09
40	17	−28	−03	22	01	−07	−14	−04	−07	−24	14	−06	09	−06	05	02	04		02	11	−06	−02	01	−02	26	−06	10	−09
44	−09	−11	19	18	−02	−00	12	−02	01	02	−06	−07	08	16	−08	07	−06	02		−31	09	−12	10	−04	−26	08	−05	16
45	09	16	13	08	−03	01	−29	14	−14	12	−04	−04	−11	05	23	−09	−14	11	−31		05	21	07	−04	−24	−06	10	−06
46	−05	−10	−16	06	05	−04	−12	08	15	−02	−04	14	−02	15	04	−06	−06	−06	09	05		−28	−11	05	−07	18	−13	30
48	−11	15	−08	−22	−01	07	−06	13	−13	07	01	00	15	−15	02	−01	22	−02	−12	21	−28		11	01	22	01	−08	−08
50	−08	−05	−28	06	25	04	−02	−16	−01	−27	−11	−12	08	08	12	26	11	01	10	07	−11	11		−03	−18	11	28	−28
53	17	−06	−12	02	01	10	−06	−03	02	−01	−07	04	15	−15	−14	13	14	−02	−04	−04	05	01	−03		03	01	−13	01
54	−02	39	−00	−26	−08	02	02	*	08	12	−05	−12	08	01	−03	04	09	26	−26	−24	−07	22	−18	03		06	−10	−03
55	*	−08	−27	−00	*	17	−05	−18	10	06	03	−05	−16	−03	19	−04	−20	−06	08	−06	18	01	11	01	06		06	12
56	−24	−10	04	02	19	−13	21	−12	03	−07	04	10	10	06	−04	12	−34	10	−05	10	−13	−08	28	−13	−10	06		06
57	07	−16	−02	−18	12	−12	−13	18	05	11	−09	10	−01	−07	08	−13	09	−09	16	−06	30	−08	−28	01	−03	12	06	

* 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

Bivariate Frequency Distribution of Ability and Score on Test 5 (Items of Medium Difficulty) Showing Frequency per 10,000; also the Marginal Frequency 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	
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 8

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

Test Score	Ability Score (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	f_s
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	...	1	11	82	309	576	548	278	80	14	2	...	...	...	...	095
6	...	3	25	125	305	369	227	73	13	1	...	...	...	...	...	057
5	1	7	41	130	208	165	65	14	1	...	...	...	...	...	...	032
4	2	11	45	93	98	51	13	2	...	...	...	...	...	...	...	016
3	3	12	32	45	31	11	2	...	...	...	...	...	...	...	...	007
2	2	8	14	13	6	1	...	...	...	...	...	...	...	...	...	002
1	1	3	3	2	1	...	...	...	...	...	...	...	...	...	...	001
0	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

TABLE 9

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

Comparison of Theoretical (italics) and Actual Statistics for the Univariate Score Distributions for the Eight Tests, Together With the Chi-Squares Between Theoretical and Actual Frequencies

Test	Kinds of Items	Chi-Square	Mean	Standard Deviation	Skewness (α_3)	Kurtosis ($\beta_2 - 3$)
2	Difficult	3.62	1.70 *1.70*	1.73 *1.65*	1.0 *1.0*	0.6 *0.6*
5	Medium Difficulty	5.51	4.21 *4.21*	2.07 *2.13*	−0.1 *0.0*	−0.8 *−0.9*
8	Easy	0.75	8.86 *8.85*	1.83 *1.83*	−1.0 *−1.0*	1.0 *0.7*
82	Easy and Difficult	3.85	5.00 *5.00*	1.81 *1.83*	0.1 *0.0*	0.0 *−0.2*
h	Highly Discriminating	18.42	3.93 *3.93*	2.44 *2.46*	0.5 *0.4*	−0.6 *−0.6*
L	Poorly Discriminating	1.99	5.43 *5.43*	1.59 *1.59*	−0.3 *−0.2*	−0.4 *−0.2*
r	Rectangular Distribution	3.07	4.47 *4.47*	1.97 *2.00*	0.1 *0.1*	−0.7 *−0.6*

TABLE 11

Scatter Diagram for Test 2 (Difficult Items) and Test 5 (Items of Medium Difficulty) Showing Actual Frequencies (integers) and Theoretical Frequencies (decimals)

Score on Test 2	Score on Test 5: 0	1	2	3	4	5	6	7	8	Total
9	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...
8	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	0.1	0.1	0.3
7	...	...	...	...	...	...	...	1	2	3
	...	...	...	...	...	...	0.1	0.4	0.5	1.1
6	...	...	...	...	...	...	1	...	1	2
	...	...	...	...	...	0.3	0.5	1.1	1.1	3.0
5	...	...	...	...	...	...	3	2	...	5
	...	...	...	0.1	0.3	0.7	1.4	1.9	1.5	5.8
4	...	...	1	...	2	...	3	2	2	10
	...	...	0.1	0.4	0.8	1.6	2.4	2.7	1.6	9.8
3	...	1	1	2	3	5	3	3	2	20
	...	0.1	0.5	1.2	2.3	3.3	3.8	3.1	1.5	15.9
2	...	2	2	2	4	4	6	...	1	21
	0.3	1.0	2.2	3.5	4.8	5.2	4.6	3.0	1.1	25.6
1	1	3	3	8	6	4	5	1	...	31
	1.2	3.4	5.7	7.1	7.1	5.8	3.9	1.9	0.5	36.7
0	4	5	6	10	9	5	4	1	...	44
	3.1	6.7	8.3	7.6	5.8	3.7	1.9	0.7	0.1	37.9
Total	5	11	13	22	24	18	25	10	8	N=136
	4.6	11.2	16.9	20.0	21.1	20.5	18.8	15.0	8.2	

TABLE 12

Scatter Diagram for Test 8 (Easy Items) and Test 2 (Difficult Items) Showing Actual Frequencies (integers) and Theoretical Frequencies (decimals)

Score on Test 8	Score on Test 2										
	0	1	2	3	4	5	6	7	8	9	Total
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 13

Scatter Diagram for Test 8 (Easy Items) and Test 5 (Items of Medium Difficulty) Showing Actual Frequencies (integers) and Theoretical Frequencies (decimals)

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

TABLE 14

Scatter Diagram for Test 8′ (Easy Items) and Test *r* (Item Difficulties Rectangularly Distributed) Showing Actual Frequencies (integers) and Theoretical Frequencies (decimals)

Score on Test *r*	Score on Test 8′ 0	1	2	3	4	5	6	7	8	9	Total
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 15

Scatter Diagram for Test 82 (Both Easy and Difficult Items) and Test 5 (Items of Medium Difficulty) Showing Actual Frequencies (integers) and Theoretical Frequencies (decimals)

Score on Test 82	Score on Test 5: 0	1	2	3	4	5	6	7	8	Total
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 16

Scatter Diagram for Test 82 (Both Easy and Difficult Items) and Test *r* (Item Difficulties Rectangularly Distributed) Showing Actual Frequencies (integers) and Theoretical Frequencies (decimals)

Score on Test 82	Score on Test *r* 0	1	2	3	4	5	6	7	8	9	Total
10	...	...	...	...	...	...	...	...	1	...	1
	...	...	...	...	...	...	...	0.1	0.3	0.1	0.5
9	...	...	...	...	...	...	...	1	2	1	4
	...	...	...	...	0.1	0.3	0.5	1.0	1.1	0.5	3.5
8	...	...	...	...	1	...	2	1	2	...	6
	...	...	...	0.1	0.4	1.1	1.8	2.3	1.9	0.5	8.2
7	...	...	...	1	...	5	4	5	1	...	16
	...	...	0.3	0.7	1.8	3.0	3.8	3.4	2.0	0.4	15.4
6	...	...	1	3	1	6	6	3	3	...	23
	...	0.3	1.0	2.4	4.4	5.7	5.3	3.5	1.5	0.3	24.3
5	...	...	4	5	10	7	5	4	...	...	35
	0.1	0.8	2.7	5.3	7.1	6.8	4.8	2.3	0.7	0.1	30.7
4	...	1	2	9	6	2	1	...	...	...	21
	0.3	1.6	4.1	6.0	6.0	4.4	2.3	0.8	0.1	...	25.6
3	1	2	5	7	4	1	...	1	...	...	21
	0.4	1.9	3.7	4.1	3.1	1.8	0.7	0.1	...	...	15.8
2	1	1	2	1	1	...	...	...	...	...	6
	0.5	1.6	2.3	1.9	1.1	0.4	0.1	...	...	...	8.0
1	...	...	1	1	...	...	...	...	...	...	2
	0.4	0.8	1.0	0.5	0.3	0.1	...	...	...	...	3.1
0	...	...	1	...	...	...	...	...	...	...	1
	0.1	0.3	0.1	0.1	...	...	...	...	...	...	.7
Total	2	4	16	27	23	21	18	15	9	1	N=136
	1.9	7.3	15.1	21.2	24.2	23.5	19.3	13.6	7.6	2.0	

TABLE 17

Scatter Diagram for Test L (Low Validity Items) and Test h (High Validity Items) Showing Actual Frequencies (integers) and Theoretical Frequencies (decimals)

Score on Test h	Score on Test L 0	1	2	3	4	5	6	7	8	9	10	Total
10	...	...	...	...	...	...	...	...	1	...	...	1
	...	...	...	...	...	0.1	0.3	0.7	0.7	0.3	...	2.0
9	...	...	...	...	...	1	2	2	...	2	...	7
	...	...	...	...	0.1	0.4	1.0	1.5	1.2	0.3	...	4.5
8	...	...	...	...	...	...	3	1	...	...	...	4
	...	...	...	...	0.3	0.8	1.8	2.3	1.5	0.3	...	6.9
7	...	...	...	...	...	3	2	8	2	...	...	15
	...	...	...	0.1	0.5	1.5	2.9	3.0	1.6	0.3	...	9.9
6	...	...	...	...	...	2	3	1	...	...	...	6
	...	...	...	0.3	1.0	2.4	3.9	3.5	1.5	0.3	...	12.9
5	...	...	...	...	3	5	6	4	2	...	...	20
	...	...	0.1	0.5	1.6	3.7	4.8	3.5	1.4	0.1	...	15.8
4	...	...	...	...	2	3	2	2	2	...	...	11
	...	...	0.3	1.0	2.7	4.8	5.3	3.3	1.1	0.1	...	18.5
3	...	...	1	6	2	8	7	4	...	...	...	28
	...	...	0.4	1.6	3.8	5.7	5.2	2.7	0.7	0.1	...	20.3
2	...	...	1	3	5	6	4	5	1	...	...	25
	...	0.1	0.8	2.6	4.9	5.7	4.2	1.8	0.4	...	...	20.5
1	...	...	1	3	4	1	1	1	...	...	...	11
	...	0.3	1.4	3.1	4.6	4.4	2.6	0.8	0.1	...	...	17.3
0	...	1	1	1	3	1	1	...	...	...	...	8
	...	0.3	1.0	1.8	1.9	1.4	0.7	0.1	...	...	...	7.1
Total	...	1	4	13	19	30	31	28	8	2	...	
	...	0.7	3.9	11.0	21.5	30.9	32.5	23.3	10.2	1.8	...	N=136

APPENDIX

SUMMARY OF NOTATION

$A(y_o) = \int_{y_o}^{\infty} N(y)dy$ = the area of the normal curve lying above any given point y_o.

$A_2(u_o,v_o;r) = \int_{u_o}^{\infty}\int_{v_o}^{\infty} N_2(u,v;r)dvdu$ (the frequency in a specified region of the bivariate normal distribution).

$B(y_o) = 1 - A(y_o)$.

c = the "underlying ability measured by the test"; the common factor of the matrix of tetrachoric item intercorrelations.

D = the discrimination index at a specified level of ability.

e = the base of the system of natural logarithms.

$\exp(y) = e^y$.

$E(y)$ = the expected or average value of any variable, y.

f_{uv} = the bivariate frequency distribution of any two variables, u and v.

f_u = the univariate frequency distribution of any variable, u.

$f_{u.v}$ = the conditional frequency distribution of any variable, u, for a fixed value of another variable, v.

F_t = the cumulative frequency distribution of t.

$g_i = \dfrac{h_i - R_i c}{K_i}$.

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; $i = 1, 2, \ldots, n$. In M_i, σ_i, r_{ic}, etc., i is used to stand for x_i.

j = subscript indicating the j-th item; $j = 1, 2, \ldots, n$. 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 i correctly (P_i is a function of c).

$q_i = 1 - p_i$.

$Q_i = 1 - P_i$.

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

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

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

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

r_{ij}' = tetrachoric correlation between items i and j.

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

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

$t = \lim_{n=\infty} z$ = 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 i; $x_i = 0$ or 1.

y = any variable.

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

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

Π_s, Π_{n-s} (see explanation of Equation 12).

$\sigma_i = \sqrt{p_i q_i}$ = the standard deviation of x_i.

σ_y = standard deviation of any variable, y.

Σ^* (see explanation of Equation 12).

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. Cronbach, L. J. and Warrington, W. G. Efficiency of multiple-choice tests as a function of spread of item difficulties. *Psychometrika,* 1952, 17.
5. Davis, F. B. Item-analysis data. Harvard University, Graduate School of Education, 1946.
6. Dunlap, J. W. and Kurtz, A. K. Handbook of statistical nomographs, tables, and formulas. Yonkers: World Book Company, 1932.
7. Ferguson, G. A. Item selection by the constant process. *Psychometrika,* 1942, 7, 19-29.
8. Ferguson, G. A. On the theory of test discrimination. *Psychometrika,* 1949, 14, 61-68.
9. Finney, D. J. Probit analysis. Cambridge: Cambridge University Press, 1947.
10. 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.
11. Guilford, J. P. Psychometric methods. New York and London: McGraw-Hill Book Company, 1936.
12. Gulliksen, H. The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika,* 1945, 10, 79-91.
13. Gulliksen, H. Theory of mental tests. New York: John Wiley and Sons; London: Chapman and Hall; 1950.
14. Hayes, S. P. Tables of the standard error of tetrachoric correlation coefficient. *Psychometrika,* 1943, **8,** 193-203.
15. Jackson, R. W. B. and Ferguson, G. A. A functional approach in test construction. *Educ. psychol. Meas.,* 1943, **3,** 23-28.

16. Kelley, T. L. Fundamentals of statistics. Cambridge: Harvard University Press, 1947, pp. 370-373.
17. Kendall, M. G. The advanced theory of statistics. London: Charles Griffin and Company, 1945.
18. Kuder, G. F. Nomograph for point biserial *r*, biserial *r*, and four-fold correlations. *Psychometrika,* 1937, 2, 135-138.
19. Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika,* 1937, 2, 151-160.
20. Lawley, D. N. On problems connected with item selection and test construction. *Proc. Roy. Soc. Edin.,* 1943, 61-A, Part 3, 273-287.
21. Lawley, D. N. The factorial analysis of multiple-item tests. *Proc. Roy. Soc. Edin.,* 1944, 62-A, Part I, 74-82.
22. 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, Chapts. 10 and 11.
23. Long, J. A. and Sandiford, P. The validation of test items. Bulletin No. 3, Department of Educational Research, University of Toronto, 1935.
24. Lorr, M. Interrelationships of number-correct and limen scores for an amount limit test. *Psychometrika,* 1944, 9, 17-30.
25. Mollenkopf, W. G. Variation of the standard error of measurement. Ph.D. thesis, Princeton University, 1948. Also *Psychometrika,* 1949, 14, 189-229.
26. Mosier, C. I. Psychophysics and mental test theory: fundamental postulates and elementary theorems. *Psychol. Rev.,* 1940, 47, 355-366.
27. Mosier, C. I. Psychophysics and mental test theory. II. The constant process. *Psychol. Rev.,* 1941, 48, 235-249.
28. Pearson, K. Mathematical contributions to the theory of evolution. *Royal Soc. London, Phil. Trans.,* Series A, 1900, 195, 1-47.
29. Pearson, K. Tables for statisticians and biometricians. London: Cambridge University Press, 1924.
30. Peters, C. and Van Voorhis, W. Statistical procedures and their mathematical bases. New York and London: McGraw-Hill Book Company, 1940.
31. Plumlee, L. B. The effect of difficulty and chance success on item-test correlation and on test reliability. *Psychometrika,* 1952, 17, 69-86.
32. Richardson, M. W. Relation between the difficulty and the differential validity of a test. Ph.D. thesis, University of Chi-

cago, 1936. Also *Psychometrika,* 1936, 1 (No. 2), 33-49.
33. Symonds, P. M. Choice of items for a test on the basis of difficulty. *J. educ. Psychol.,* 1928, 19, 73-87.
34. Thorndike, R. L. Personnel selection. New York: John Wiley and Sons, 1949, pp. 228-230.
35. Thurstone, T. G. The difficulty of a test and its diagnostic value. *J. educ. Psychol.,* 1932, 23, 335-43.
36. Tucker, L. R. Maximum validity of a test with equivalent items. *Psychometrika,* 1946, 11, 1-13.
37. Tucker, L. R. A method for scaling ability test items in difficulty taking item unreliability into account. *Amer. Psychologist,* 1948, 3, 309-10. (Abstract.)
38. Wherry, R. J. and Gaylord, R. H. Factor pattern of test items and tests as a function of the correlation coefficient: content, difficulty, and constant error factors. *Psychometrika,* 1944, 9, 237-44.
39. Wilks, S. S. Mathematical statistics. Princeton: Princeton University Press, 1944.