PSYCHOMETRIKA—VOL. 30, NO. 3
SEPTEMBER, 1965

A STRONG TRUE-SCORE THEORY, WITH APPLICATIONS*

FREDERIC M. LORD

EDUCATIONAL TESTING SERVICE
AND
PRINCETON UNIVERSITY

A "strong" mathematical model for the relation between observed scores and true scores is developed. This model can be used

1. To estimate the frequency distribution of observed scores that will result when a given test is lengthened.
2. To equate true scores on two tests by the equipercentile method.
3. To estimate the frequencies in the scatterplot between two parallel (nonparallel) tests of the same psychological trait, using only the information in a (the) marginal distribution(s).
4. To estimate the frequency distribution of a test for a group that has taken only a short form of the test (this is useful for obtaining norms).
5. To estimate the effects of selecting individuals on a fallible measure.
6. To effect matching of groups with respect to true score when only a fallible measure is available.
7. To investigate whether two tests really measure the same psychological function when they have a nonlinear relationship.
8. To describe and evaluate the properties of a specific test considered as a measuring instrument.

The model has been tested empirically, using it to estimate bivariate distributions from univariate distributions, with good results, as checked by chi-square tests.

True-score theory attempts to provide a mathematical model for the relation between obtained fallible measurements (test scores) and the error-free measurements that one would prefer to obtain. A successful true-score theory predicts mental-test results before they have been observed. To use a currently popular term with significant implications, a successful mathematical model *simulates* (in part) the test-responding behavior of individuals or groups of examinees.

Granted that parallel test forms can be constructed, the true-score theory in Gulliksen's early chapters ([9], chs. 2–9 excepting sect. 2.11) cannot be contradicted by any set of test data whatsoever. The true-score theory

*This work was supported in part by contract Nonr–2752(00) between the Office of Naval Research and Educational Testing Service. Reproduction in whole or in part for any purpose of the United States Government is permitted. Many of the extensive computations were done on Princeton University computer facilities, supported in part by National Science Foundation Grant NSF–GP579. The last portion of the work was carried out while the writer was Brittingham Visiting Professor at the University of Wisconsin. The writer is indebted to Diana Lees, who wrote many of the computer programs, checked most of the mathematical derivations, and gave other invaluable assistance throughout the project.

in Lord [18] is also a "weak," distribution-free model incapable of being contradicted by actual data. The model to be outlined here is a "strong" model, relying on assumptions about the distribution of the errors of measurement; in the form treated in detail here, assumptions are also made about the mathematical form of the true-score distribution. Both strong and weak models are useful: the strong models, because when successful they permit stronger inferences than can be obtained from the weak models; the weak models, because they are likely to be successful when the strong models fail.

The strong model outlined in sections 1 and 2 has proven to be a successful one for a number of types of data, as shown by the results summarized and discussed in sections 5 and 6. Practical implications of the model are outlined in section 4; one empirical application (a "selection problem") is briefly illustrated in section 7. Section 8 derives the necessary formulas.

The results reported in section 6 appear to the writer to establish the utility and importance of the true-score model used. It is possible that with the aid of electronic computers the approach outlined can, with modifications, eventually become a normal part of practical test evaluation and interpretation.

1. *The Basic Mathematical Model*

Let $\phi(x)$ denote the frequency distribution of the observed score, x, in a population of examinees; let $g(\zeta)$ be the distribution in the same population of some hypothetical ability (or other trait) measured by the test; and let $h(x \mid \zeta)$ be the conditional distribution of x for a given value of ζ. A perfectly general formula states that

$$\phi(x) = \int_{-\infty}^{\infty} g(\zeta)h(x \mid \zeta)\, d\zeta, \tag{1}$$

provided the integration is appropriately defined.

Actually (1) is likely to be a basic equation for any science that must make inferences from fallible measurements. The general branch of scientific inference that concerns itself with (1) will be called *latent variable theory*; in the specific applications considered here, it will be called *latent trait theory.*

If $h(x \mid \zeta)$ is sufficiently well specified, the integral equation (1) may in many cases be solved, thus determining from $\phi(x)$ a possible mathematical form for $g(\zeta)$ (see [26]—note that x is here a discrete and bounded variable; any solution valid for continuous x is also valid for discrete x, but the solution is not unique). In practice, however, the population distribution $\phi(x)$ is not available. Instead, the practical worker has only the distribution $f(x)$, say, for a sample of N examinees. Although $f(x)$ provides an estimate of $\phi(x)$, it is unwise ([19], p.4) to apply methods for solving exact integral equations to the inexact result of substituting $f(x)$ for $\phi(x)$ in (1).

Various approaches to estimating $g(\zeta)$ from $f(x)$ are available in the

mathematical literature (e.g., [2]; [3]; [12]; [15]; [21]; [27], chs. 1.4 and 1.5; [31]). (Also see May 1960 *Mathematical Reviews* for many references). The use of a quadrature formula [see 8, 11, 32, 33] is a favored standard procedure. After laborious trials of several methods, however, the writer concluded that for present purposes most such formulas either provide too poor a fit to the integrand for extreme values of x; or else have too many parameters that must be estimated from the data, leading to undesirable irregularities in the estimated $g(\zeta)$. The quadrature methods of Phillips [22] and of Twomey [29], designed to minimize these irregularities, help to overcome these objections.

The method used in the present paper for estimating $g(\zeta)$ from $f(x)$ is to assume a functional form for $g(\zeta)$ with free parameters and then to determine these parameters by use of (1).

All approaches to using (1) for estimating $g(\zeta)$ require that $h(x \mid \zeta)$ be of a known form. If the observed test score, x, is a continuous variable assuming all values from $-\infty$ to $+\infty$, then it may be appropriate to assume that $h(x \mid \zeta)$ is some specified normal distribution, in which case many very strong conclusions can be drawn (e.g., [10]; [17], pp. 292-296; [23]; [25]; [27], chs. 1.4 and 1.5).

Here, on the contrary, we shall consider only the most usual case where test score x is a positive integer, $0 \leq x \leq n$. In this case, x will usually be simply the number of right answers on a test composed of n test "questions" or "items"; for convenience, the exposition will be written as if this were always so.

Clearly, in this case, $h(x \mid \zeta)$ cannot be a normal distribution. The choice of functional forms to represent $h(x \mid \zeta)$ and $g(\zeta)$ of course depends on what sort of latent variable, ζ, is to be investigated. Here, we wish the difference $x/n - \zeta$ to have the properties usually ascribed in mental test theory to errors of measurement (e.g., [9], chs. 2, 3). If we are successful, then. since x/n is the proportion of test items answered correctly, ζ will have the properties usually ascribed to the "true (proportion-correct) score," here called simply the *true score*. It follows that $0 \leq \zeta \leq 1$.

With this in view, it will be assumed for present purposes that $h(x \mid \zeta)$ is a compound binomial distribution (see, for example, [14], vol. 1, sect. 5.10). A more detailed specification of $h(x \mid \zeta)$ is given in the following section and in section 8.

2. *A Specific Mathematical Model*

It is here assumed that $g(\zeta)$ is a four-parameter (incomplete) beta distribution:

$$(2) \qquad g(\zeta) = \begin{cases} \dfrac{(-a + \zeta)^{\alpha-1}(b - \zeta)^{\beta-1}}{A^{\alpha+\beta-1}B(\alpha, \beta)}, & \text{for } 0 \leq a \leq \zeta \leq b \leq 1, \\ 0, & \text{for } \zeta < a \text{ or } \zeta > b, \end{cases}$$

where

$$A = b - a, \tag{3}$$

and

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{4}$$

is the usual (complete) beta function. Equation (2) defines a frequency distribution only when $\alpha, \beta > 0$. Since ζ is thought of as a proportion of items answered correctly, it follows that here $0 \leq a \leq \zeta \leq b \leq 1$.

Reasons for choosing (2) include the following.

1. Surprisingly good results (although less than completely satisfactory) had been obtained with a two-parameter beta distribution [13].

2. The use of a $g(\zeta)$ with more than four parameters was found [e.g., 19] to give irregularly shaped distributions or "distributions" with negative frequencies in one or both tails.

3. Empirical investigations [19], and also theoretical considerations, showed the need of using a form for $g(\zeta)$ permitting a nonzero lower bound for ζ. One reason for this requirement is that examinees who answer all test items at random necessarily have a true score of $\zeta = 1/K$, where K is the number of choices per item.

4. Equation (2) was found to provide good fits when applied to a variety of test data, as will be shown in sections 5 and 6.

As previously noted, it is assumed that $h(x \mid \zeta)$ is a compound binomial. In practice, a crude approximation, called $P(x \mid \zeta)$, actually is used here instead:

$$h(x \mid \zeta) = P(x \mid \zeta) = p_n(x) + k\zeta(1 - \zeta)C_2(x), \qquad (x = 0, 1, \cdots, n), \tag{5}$$

where

$$p_n(x) = \begin{cases} \binom{n}{x}\zeta^x(1 - \zeta)^{n-x}, & \text{for} \quad x = 0, 1, \cdots, n, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

$$C_r(x) = \sum_{u=0}^{r} (-1)^{u+1}\binom{r}{u}p_{n-r}(x - u), \qquad (r = 2, 3, \cdots, n), \tag{7}$$

and k is a parameter of the distribution. A method for estimating k is outlined in the following section.

Equation (5) is actually a specialization of the first two terms of an n-variable Taylor-series expansion of the compound binomial distribution [see 30]. This somewhat crude approximation is defended not on theoretical grounds, but on the grounds that it seems to work. In fact, using just the first term of (5) was found to give good, although not entirely satisfactory,

results as will be seen in sections 5 and 6. (Formulas using the first four terms of the Taylor series are being used in current investigations.)

$P(x \mid \zeta)$ is actually a frequency distribution for given ζ only when k is small enough so that $P(x \mid \zeta) \geq 0$ for $x = 0, 1, \cdots, n$. The precise conditions under which this holds will be discussed in section 8; in the meantime, to avoid awkwardness and repetitiousness, $P(x \mid \zeta)$ will be spoken of as a frequency distribution, or at least as a close approximation to one, as indeed it appears to be for all relevant cases of practical importance so far studied.

On the basis of (2) and (5), (1) becomes, for present purposes,

$$\phi(x) = \int_a^b \frac{(-a + \zeta)^{\alpha-1}(b - \zeta)^{\beta-1}}{A^{\alpha+\beta-1}B(\alpha, \beta)} \left[\binom{n}{x} \zeta^x (1 - \zeta)^{n-x} + k\zeta(1 - \zeta)C_2(x) \right] d\zeta, \qquad (x = 0, 1, \cdots, n), \tag{8}$$

where $0 \leq a < b \leq 1$, $\alpha > 0$, $\beta > 0$, $k \geq 0$.

3. *Current Implementation of the Model*

The model defined by (8) has six parameters: n, a, b, α, β, and k. One of these, n, the number of items in the test, does not need to be estimated from the data since it is known exactly in advance of any testing. The others must be estimated. Currently, applications of this model are based on samples of 1000 or more, preferably 2000 or more, examinees. Since the samples are large, sample statistics will in most cases be used in lieu of the corresponding population parameters. Once the utility of the model has been shown, further consideration can be given to the efficiency of the estimates and to their sampling characteristics.

The value of k is currently determined so that the correlation between x and ζ under the model will be equal to the square root of the Kuder-Richardson formula-20 reliability coefficient (r_{20}) computed from the sample of examinees at hand. This result is achieved (see section 8) by setting

$$k = \frac{n^2(n - 1)s_p^2}{2[\bar{x}(n - \bar{x}) - s_x^2 - ns_p^2]}, \tag{9}$$

where $\bar{x}$ and s_x^2 are the mean and variance of $f(x)$, the actual score distribution, and s_p^2 is the variance of the usual item difficulties. Appreciable fluctuations in s_p^2 produce rather small fluctuations in r_{20} (and very small changes in $\phi(x)$); so s_p^2 may sometimes be estimated from an ordinary item analysis based on a relatively small sample of the answer sheets, or it might even be estimated subjectively by armchair procedures.

It will be indicated in section 8 that k can be expected to remain reasonably constant from one group of examinees to another, assuming the groups to be sufficiently similar so that the trait measured by the test remains the same for all groups.

It will be seen that the main importance of k is its effect on the estimated product-moment correlation for two parallel tests. For the present, r_{20} is taken as a sufficiently good estimate of this correlation. If a better estimate of the parallel-forms correlation were available, it could be substituted for r_{20} and a correspondingly better value of k determined by the line of reasoning given in section 8.

The parameters a, b, α, and β are here estimated by the method of moments, so that the $\phi(x)$ computed from (8) will have the same first four moments as the actual $f(x)$. Since only four moments of $f(x)$ are used to fit the model, a chi-square test between $f(x)$ and $\phi(x)$, to check on the model, has $n - 4$ degrees of freedom provided the distributions are ungrouped.

The fact that this chi-square has $n - 4$ degrees of freedom suggests an otherwise obvious conclusion: that $g(\zeta)$ is ordinarily not uniquely determined when $n < 4$. It is thus to be assumed throughout that $n \geq 4$. In empirical applications to date, ordinarily $n \geq 25$.

4. *Some Implications and Applications of the Model*

This section is concerned with both theoretical implications and practical applications of the model. It is understood that in practical applications, first of all the parameters of the model are to be estimated from the data. It is assumed that the number of examinees is sufficiently large so that useful results can be obtained by substituting estimates of parameters for their true values. When necessary for clarity, a "hat" ($\hat{}$) will be placed on a symbol to indicate that it refers to an estimated quantity.

Distribution of Observed Scores

Clearly, the model implies that when n and appropriate estimates of the remaining five parameters are inserted, the $\hat{\phi}(x)$ calculated by carrying out the integration in (8) should provide a good fit to the distribution of observed test scores. This has been checked by chi-square tests on a variety of data, as reported in section 5. The $\hat{\phi}(x)$ thus obtained is called the *fitted distribution.*

Relation of Observed Score to True Score

Once adequate estimates have been obtained for the parameters of the model, the estimated bivariate frequency distribution of observed score and true score can be computed by substituting estimates into the formula

$$F(x, \zeta) = g(\zeta)h(x \mid \zeta). \tag{10}$$

$F(x, \zeta)$ provides complete information about the relation of x to ζ. In particular, the (curvilinear) regression of ζ on x can be estimated from (10), as can the scatter about this regression. This result permits estimation of ζ for an examinee from a knowledge of his observed x.

A further application of (10) is outlined later in the present section (*Some Selection Problems*). Other applications for this phase of the present study will be discussed in another publication.

Lengthening the Test

The effect of lengthening (or shortening) a given test by adding suitably "parallel" forms may be found by changing n in the formula of interest. Effects on the shape of the observed score distribution, for example, can be determined from (8). It may be shown, for example, that the skewness of the observed-score distribution may either increase or decrease as the test is lengthened, depending on the specific values of the six parameters. The equations obtained from the model go beyond, but are not inconsistent with, formulas already available in test theory ([9], chs. 5–9)—for example, the formulas for the effect of lengthening a test on the mean and variance of its observed-score distribution, on the mean and variance of the errors of measurement, and on the correlation between observed score and true score.

A special application of this approach is mentioned later in this section (see *The Norms Problem*).

Equating

If it is stated that two different tests measure the same psychological trait, this may be interpreted as meaning that their true scores will have a perfect functional relation for any group of examinees. Let ζ and η be the true scores on the two tests, and (for present purposes) let the functional relation $\eta = \psi(\zeta)$ be strictly increasing. Suppose both tests have been administered to the same group of examinees; let g and G be the two true-score distributions, respectively. Then $\psi(\zeta)$ can be determined from

$$\int_0^{\zeta_0} g(\zeta)\, d\zeta = \int_0^{\psi(\zeta_0)} G(\eta)\, d\eta. \tag{11}$$

Equation (11) states a relationship that is familiar in equipercentile equating.

In the course of other work (see section 6), (11) has now been solved numerically for many pairs of tests, an estimated value of η being determined for each of many specified values of ζ, for each pair of tests (for example, see Figs. 7 and 8).

Estimating the Shape of Scatterplots

If the errors of measurement in x are unrelated to those in y, as will here be assumed, then the conditional joint distribution (H_2 , say) of x and y may be written

$$H_2(x, y \mid \zeta) = H_2(x, y \mid \eta) = h(x \mid \zeta)H(y \mid \eta), \tag{12}$$

where $H(y \mid \eta)$ is the conditional distribution of y for given η, here represented

by (5) with x replaced by y, ζ by η, and n by m, the number of items in test y. If $\phi(x, y)$ is the bivariate distribution of the observed scores on two tests measuring the same psychological trait, then

$$\phi(x, y) = \int_0^1 g(\zeta)h(x \mid \zeta)H(y \mid \eta)\, d\zeta, \tag{13}$$

$$(x = 0, 1, \cdots, n; \qquad y = 0, 1, \cdots, m),$$

where $\eta = \psi(\zeta)$, as required by (11). Equation (13) may be viewed simply as an additional part of the mathematical model under consideration.

When the six parameters for each test have been determined, (11) can be solved numerically to determine an estimated η for any given ζ; (13) can then be integrated numerically to determine an estimate, $\hat{\phi}(x, y)$, for each value of x and of y. Thus the scatterplot between two such tests can be estimated from the two univariate distributions. If x and y are strictly parallel, then $\eta \equiv \zeta$ and the bivariate distribution of x and y can be estimated from the univariate distribution of either test.

Equation (13) provides a much more crucial empirical test of the predictive value of the model than does (8). Consequently (13) has been computed for quite a number of pairs of tests so that the resulting $\hat{\phi}(x, y)$ could be compared to the actually observed $f(x, y)$. The results are reported in section 6.

The Norms Problem

Publishers of grade school and high school tests usually try to provide representative national norms. Unfortunately, this is ordinarily quite impossible, since a test publisher can only test in the more or less select group of schools that are willing to cooperate with him.

Occasionally there may be nationwide efforts, perhaps sponsored by the National Science Foundation, the U. S. Office of Education, or other prestigious groups, which succeed in obtaining cooperation from a truly representative nationwide sample of schools. Whereas it might be out of the question to administer many long tests in this sample of schools, it might be possible to administer several very short tests requiring only a few minutes each. The "norms problem" then is to estimate effectively from data on a short version how the nationwide sample would have performed if the complete test could have been administered. Surely an adequate theory of mental tests should provide a method of answering such a question.

If the short test and the long test are strictly parallel except for their different lengths, then they will have the same $g(\zeta)$. In this case, it would only be necessary to estimate the five parameters other than n from the data on the short test and substitute these estimates in (8), using an n equal to the number of items in the long test. The $\hat{\phi}(x)$ computed from (8) in this way would be the (estimated) national norms distribution for the long test.

If the short test and the long test are not strictly parallel except for

length but do measure the same psychological dimension, then a more complicated procedure will be appropriate. In addition to testing the nationwide sample with the short test, the publisher should also arrange to test a large number of examinees who at least cover the range of ζ found in the nationwide sample. Both short and long tests should be administered to the latter group, which will be called the publisher's sample. To avoid practice effect, however, the short test could be administered to one random half of the publisher's sample and the long test to the other half. The statistical analysis consists of the following steps.

1. The four parameters of $g(\zeta)$ for the nationwide administration are estimated for the short test; $g(\zeta)$ is computed for numerous equally spaced values of ζ.

2. True score on the long test (η) is equated to true score on the short test (ζ), by applying (11) to the publisher's sample. Specifically, $\psi(\zeta)$ is evaluated for each value of ζ obtained in step 1.

3. The function $H(y \mid \eta)$ is evaluated for $y = 0, 1, \cdots, m$, for each value of η obtained in step 2. The value of k used is estimated from the publisher's sample (reasons for expecting k to remain roughly constant from one group to another are given in section 8).

4. According to the model, the desired national norms distribution for the long test is given by

$$\phi(y) = \int_0^1 g(\zeta)H(y \mid \eta)\, d\zeta, \qquad (y = 0, 1, \cdots, m), \tag{14}$$

where all distributions are for the nationwide sample. Estimated values of the integrand on the right of (14) are obtained using the estimated values of $H(y \mid \eta)$ computed in step 3 and estimated values of $g(\zeta)$ computed in step 1. The integral is then evaluated by numerical quadrature.

If a publisher finds experimentally that the present true-score model can accurately predict the scatterplot between the short test and the long test from their univariate distributions, then he should be well satisfied with the estimated national norms obtained by the foregoing procedure.

Of course, the "norms problem" is a more general problem than the name would indicate. A more general formulation would be as follows. Suppose tests x and y, both measures of the same trait, are administered to different random halves of Group I; and suppose that only test x is administered to Group II. The problem is to estimate the distribution of observed scores that would have been found for test y if it had been administered to Group II. The bivariate distribution of tests x and y for Group II can be estimated also, if desired.

Some Selection Problems

There are many selection problems. Three examples will be mentioned here.

Suppose a college admits all applicants with scores $x \geq x_0$ on test x. First, what is the shape of the true-score distribution of the admitted students? If x_0 is above the mode, the observed-score distribution of the admitted students is a J-shaped distribution with high positive skewness. But is their true-score distribution positively or negatively skewed? As a matter of fact, it may be either. The exact shape can be determined by estimating $g(\zeta)$ for the total group of applicants, and computing with the aid of (10), for selected values of ζ, the function $\sum_{x=x_0}^{n} F(x, \zeta)$. The computed values thus obtained are the estimated ordinates for the frequency distribution of ζ in the admitted group.

The answers obtained in the preceding paragraph are of interest, but they are not readily subjected to empirical check. If one wishes to check the model, one can ask what will be the distribution of observed scores of the admitted group on a test form y parallel to test x. This question can be answered by using (13) to obtain $\hat{\phi}(x, y)$. The estimated distribution of y in the admitted group is then simply

$$\sum_{x=x_0}^{n} \hat{\phi}(x, y), \qquad (y = 0, 1, \cdots, n).$$

A numerical illustration is given in section 7.

A third selection problem is the following. Test x has been administered to two rather different groups with N_1 and N_2 examinees, respectively; the problem is to select from Group II, solely on the basis of the test x scores, N_3 examinees such that the selected subgroup will exactly match Group I on the trait measured by test x. The desired result will be achieved if we imagine a test y strictly parallel to test x and select a subgroup of examinees from Group II so that $f_3(y)$, their hypothesized observed-score distribution on test y, is equal to $f_1(x)$, the Group I observed-score distribution on test x. The detailed procedures for solving this problem can be deduced by the reader.

A Basic Problem of Measurement

A very basic problem in any field of measurement is to be able to determine whether or not two instruments are measuring the same thing except for errors of measurement. To date, there has been no way to answer this question in mental test theory without assuming a linear relation between the tests. If the model summarized by (8) holds for different forms of test x considered separately, and also for different forms of test y, then a significant chi-square between $\hat{\phi}(x, y)$ and $f(x, y)$, as mentioned in connection with (13), is sufficient to reject the null hypothesis that tests x and y are measuring the same trait except for errors of measurement. If the chi-square is non-significant, the null hypothesis can be retained.

In practice, this significance test will lead to rejecting the null hypothesis

somewhat more often than the nominal significance level would indicate—first, because the methods used for estimating the parameters needed in (13) are less than optimal, thus failing to minimize the obtained chi-square; second, because any inadequacies of the model itself also tend to increase the size of the obtained chi-square.

5. *Empirical Results for Univariate Distributions*

Chi-Squares

Equation (8) has been used to obtain $\hat{\phi}(x)$, here called the "fitted distribution," for 16 different observed-score distributions representing nine different tests, two different scoring methods, and five different groups of examinees. Only one of the 16 chi-squares was significant at the 5 per cent level. Just eight of the 16 were significant at the 50 per cent level. (The necessary computations were carried out on an IBM 7090—see section 8.)

The size of the groups studied is, of course, crucial for interpreting statistical significance in terms of practical significance. Statistical significance here is of interest only as an index of practical significance, since the model could surely be proven false if it were tested on a sufficiently large group of examinees. The N's for the five groups were 1000, 2000, 2385, 2523, and 6113. The largest group was the one for which the significant chi-square was found. (Some groups with more than 100,000 examinees have also been investigated; however, significant chi-squares were expected for groups of this size.)

It should be noted that no grouping of frequencies was done in computing the chi-squares except as was necessary to secure theoretical frequencies of 1.0 or more in each class interval. This procedure, advised by Cochran [4], is more stringent than is the usual coarser grouping, quickly disclosing discrepancies in the tails of the distributions that would otherwise escape detection.

Illustrative Results

Since the chi-squares were so satisfactory, let us go on to look at some of the results. Figs. 1, 2, and 3 each show an actual observed-score distribution, the corresponding fitted distribution for observed scores obtained from (8), and the corresponding estimated distribution for true scores. All three tests consist of five-choice vocabulary items administered to a nationwide sample of college seniors. Test P contains 60 representative items from a longer test; Test J contains 50 mostly very easy items from the same test; Test H contains 25 mostly very hard items. The figures show that no examinees have estimated true scores below the chance level, .20, this being the proportion of successes to be expected in random guessing on five-choice items.

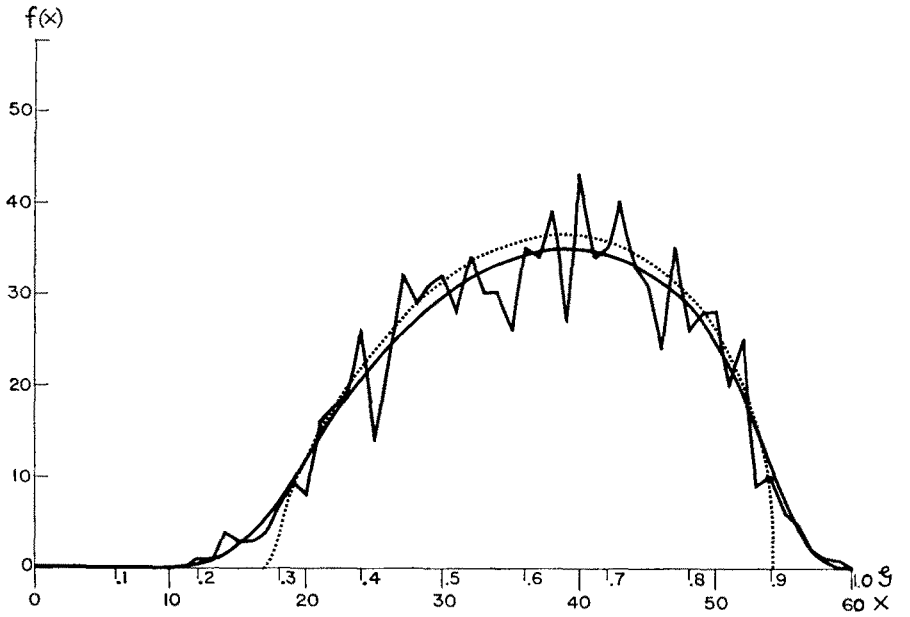

FIGURE 1
Observed-score distribution (irregular polygon), estimated true-score distribution (dotted curve), and fitted distribution, for Test *P*

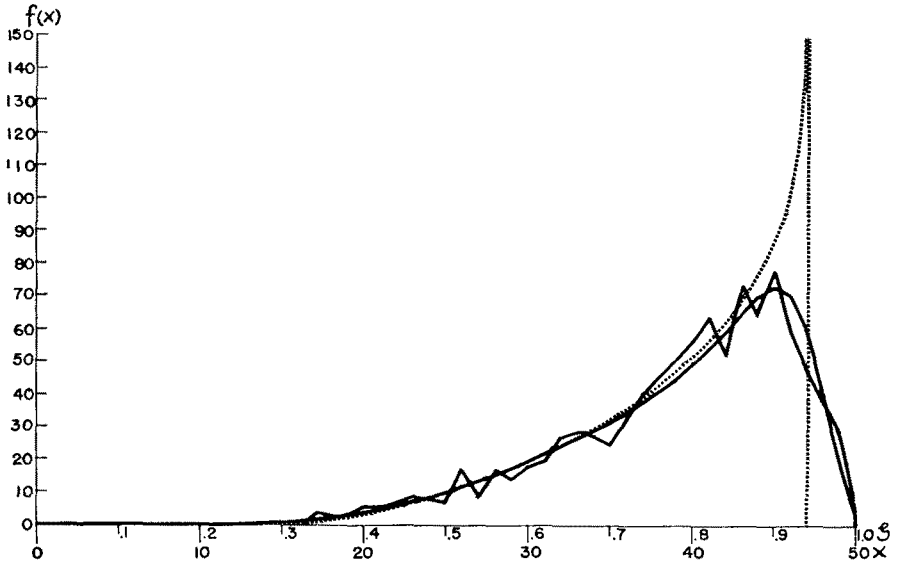

FIGURE 2
Observed-score distribution (irregular polygon), estimated true-score distribution (dotted curve), and fitted distribution, for Test *J*

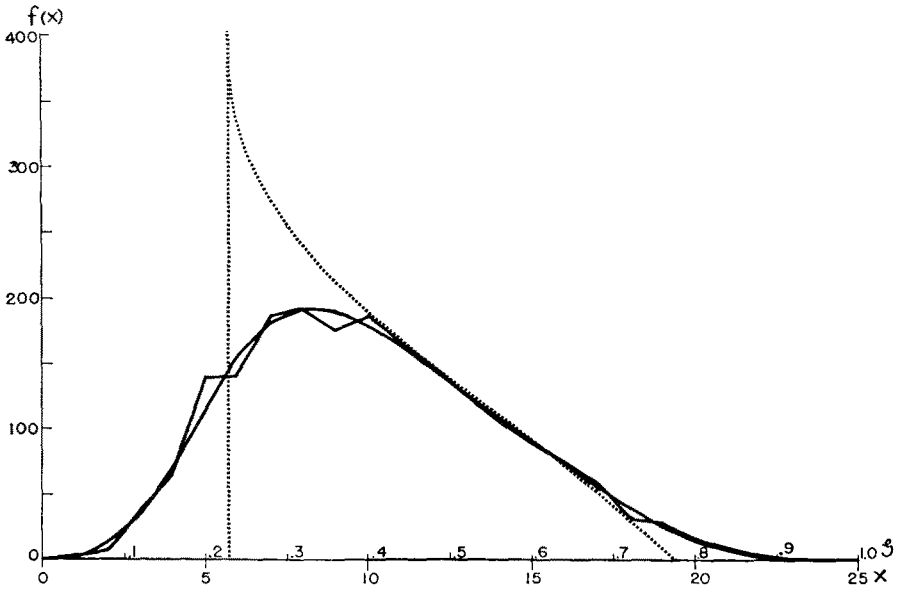

FIGURE 3
Observed-score distribution (irregular polygon), estimated true-score distribution (dotted curve), and fitted distribution, for Test *H*

On Tests *P* and *J*, and on most of the others that are not shown, the lower end point of the true-score distribution is quite inaccurately estimated since there are few people near to it. On this account, Test *H* and another test, ETSCN, shown in Fig. 4, were specially chosen for study since many people scored near the chance level on these two tests. Test ETSCN is a nationally used test, containing 30 five-choice nonverbal-reasoning items. The data represent a particular intact group of 2385 examinees encountered in one of the regular testings for whom the test was very difficult. The chi-square for ETSCN, like the others, was not significant.

It is quite clear from Figs. 3 and 4 that the effective range of the true-score distribution does not extend all the way down to $\zeta = 0$. On Test *H* the estimated lower end of the range was $\zeta = .23$; on ETSCN, it was .15. If we force the lower end of the range on ETSCN up to $\zeta = .20$, the chi-square becomes highly significant. Although too much confidence cannot be placed on exact numerical values, the results for ETSCN suggest that on this test many of the examinees may possibly be performing consistently more poorly than if they had guessed at random. This result could occur because of the attractive "distracters" ingeniously provided by the item writers.

Figs. 1 through 4 illustrate some of the more interesting shapes assumed by the estimated true-score distributions. It is clear that a casual glance

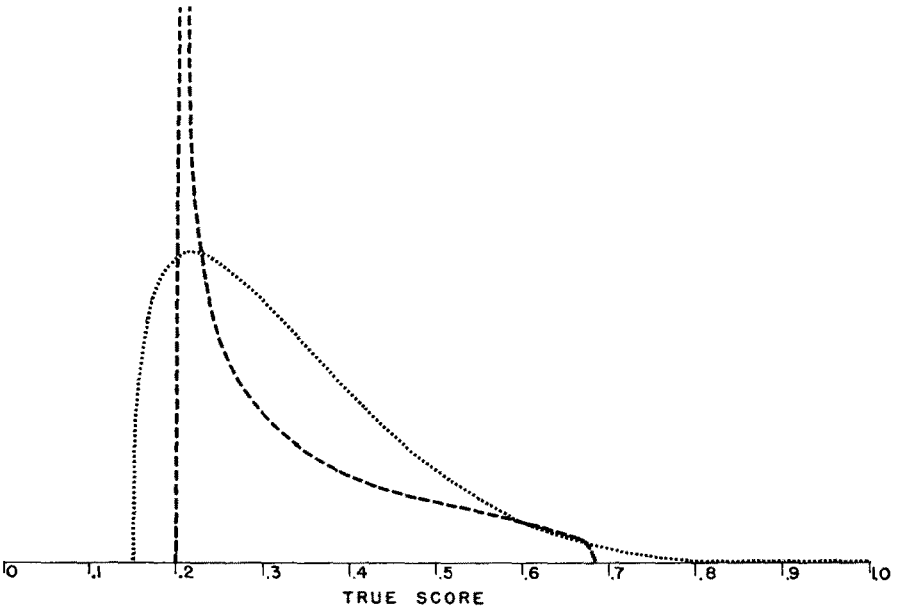

FIGURE 4
Estimated true-score distribution for ETSCN (dotted line); also the same distribution when lower end is required to be 0.2 (dashed line).

at the observed-score distribution is not always a sufficient basis for inferring the shape of the true-score distribution.

It should be remembered that proportion-correct true score is necessarily bounded between 0 and 1; such a variable will not always have the familiar bell-shaped type of distribution (bell-shaped distributions of true scores were obtained, but did not seem of sufficient novelty to require an additional figure to illustrate them here). On the other hand, too much emphasis should not be placed on the details of the shape of a particular estimated distribution—even a distribution with a vertical asymptote at one end can usually be changed to a two-tailed distribution by a rather minor modification.

The fact that estimated true scores are rarely found below .15 or .20 is an outcome of the analysis, not a requirement of the model. Since this is what would be expected for five-choice items when examinees have few omissions, this outcome increases our confidence in the meaningfulness of the estimates obtained.

Simplified Model

In fitting the 16 distributions mentioned above, the value of k used in (8) was determined from (9). Thirteen of these distributions, and many others, have also been fitted by using (8) with k arbitrarily set equal to zero

[20]. In general, the results are almost indistinguishable, as may be seen from the following list, which shows for each distribution and for each method the probability that the obtained chi-square would be exceeded by chance alone.

	P	J	H	ETSCN	S	T	U	Q	MACAA	H'	J'	P'	Q'
$k \neq 0$:	.29	.11	.64	.13	.68	.87	.58	.40	.005	.62	.68	.19	.54
$k = 0$:	.28	.13	.63	.17	.68	.90	.61	.45	.02	.60	.82	.24	.60

The most apparent systematic difference between the two methods is that the range of $\hat{g}(\zeta)$ is slightly larger when $k \neq 0$ than when $k = 0$. The following list gives the lower limit of $\hat{g}(\zeta)$ for the tests shown in Figs. 1 to 4, as found by the two methods.

	P	J	H	ETSCN
$k \neq 0$:	.264	.267	.225	.140
$k = 0$:	.276	.273	.229	.150

Figs. 1 to 4 actually show the results obtained with $k = 0$—there was no need to redraw the figures to make almost invisible changes.

In view of the similarities just described, why bother with $k \neq 0$? Because the ability to fit univariate observed-score distributions is not a very searching requirement for a true-score model. The results reported in section 6 will show why the use of $k = 0$ is inadequate.

The Problem of Omitted Responses

The mathematical models discussed so far have all been for dichotomously scored items. The logic underlying the models makes no provision for the fact that at least three and perhaps four different types of item responses can be distinguished on a typical answer sheet: right, wrong, "skipped," and "not reached."

Theoretically, (5) cannot be adequate for examinees who omit many items. For this reason, 12 of the 16 observed-score distributions just discussed were obtained by a somewhat unconventional test-scoring procedure: the scoring machine supplied a random response for each item omitted by each examinee. The result is the same as if each examinee had been required to choose *at random* among the five item responses instead of omitting an item. This type of score will be called an R-plus-r-score to distinguish it from the usual "R-score," which is simply the number of right answers.

To date both R- and R-plus-r-scores have been obtained for each of 12 tests, also chi-squares evaluating the fit of the model under the two methods of scoring. The chi-squares for one method of scoring are not highly correlated with those for the other; six of the 12 chi-squares favor the R-scores and six favor the R-plus-r-scores. It thus appears possible that the model

is sufficiently robust so that R-plus-r-scoring or other special treatment of omits may not often be necessary.

6. *Empirical Results for Bivariate Distributions*

Equations (11) and (13) have so far been used to predict 16 different bivariate frequency distributions involving three different groups of examinees

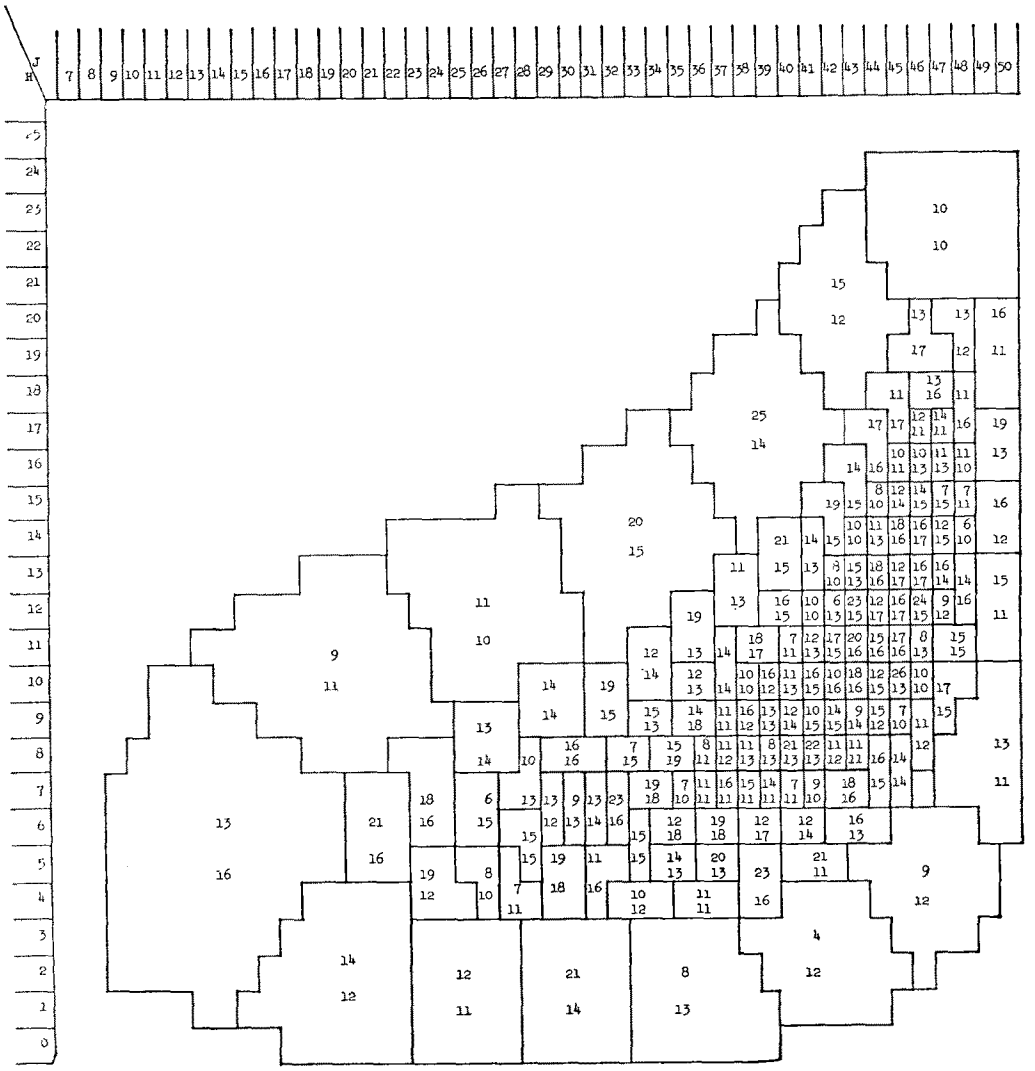

FIGURE 5
Actual frequencies (upper) and predicted frequencies (lower) for Tests H and J

and eight different vocabulary tests scored by two different methods. The N's for the three groups were 1000, 2000, and 2523.

For one pair of tests, H and J, the four chi-squares obtained were all significant at the 5 per cent level. The writer's conclusion is that a difficult vocabulary test like H, which uses such unusual key words as *limnetic, eclogue, newel, sericeous,* measures something slightly different from an easy vocabulary test such as J, which includes such key words as *renegade, clemency, irritability.* This viewpoint, persuasive by itself, tends to be substantiated by the fact that for these four bivariate distributions the observed product-moment correlation was from .02 to .05 lower than the predicted correlation, whereas for the remaining 10 bivariate distributions the observed correlation was in every case a trifle higher than the predicted correlation. A list of some of these correlations is given below in another connection.

Fig. 5 compares predicted and observed bivariate distributions ($N = 2000$) for Tests H and J, the hard and easy vocabulary tests. Fig. 6 shows for the same data the theoretical regressions of J on H and H on J; also those row means and column means of the observed distribution based on five or more cases. To the naked eye, the fit in these two figures seems rather good; the chi-square is significant at the 0.5 per cent level, however. If the two tests measure slightly different psychological traits, as suggested above, then significant chi-squares are to be expected. The analysis carried out

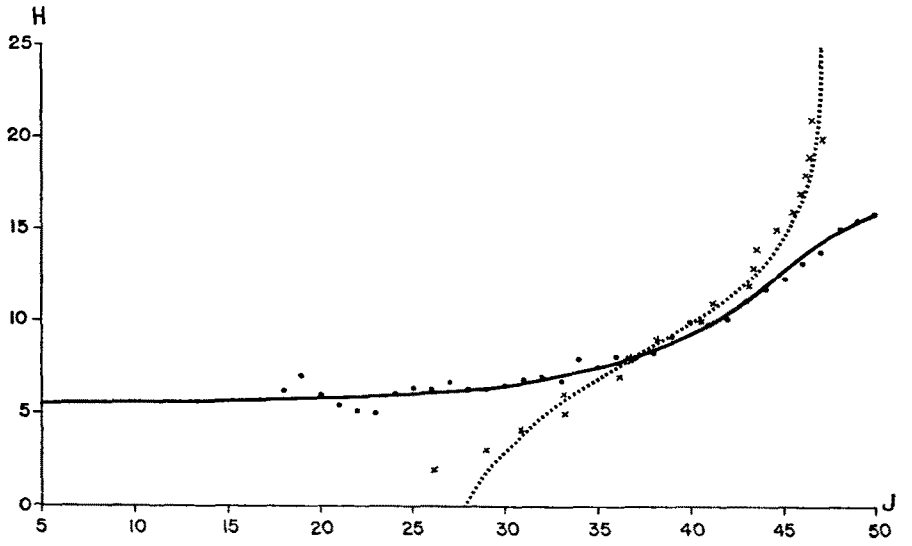

FIGURE 6
Theoretical regression of H on J (solid line) and J on H (dotted line) with actual column means (dots) and actual row means (crosses)

is in fact just the analysis that could be used to investigate whether the tests actually are or are not measuring the same dimension.

For the remaining 12 pairs of distributions studied, it is plausible to assume that both tests are measures of the same trait. For these, the model appears to be very effective: 11 of the 12 chi-squares are nonsignificant at the 5 per cent level.

A comparison of the results for the two different scoring methods shows three comparisons favoring the R-plus-r-scores and two favoring the R-scores.

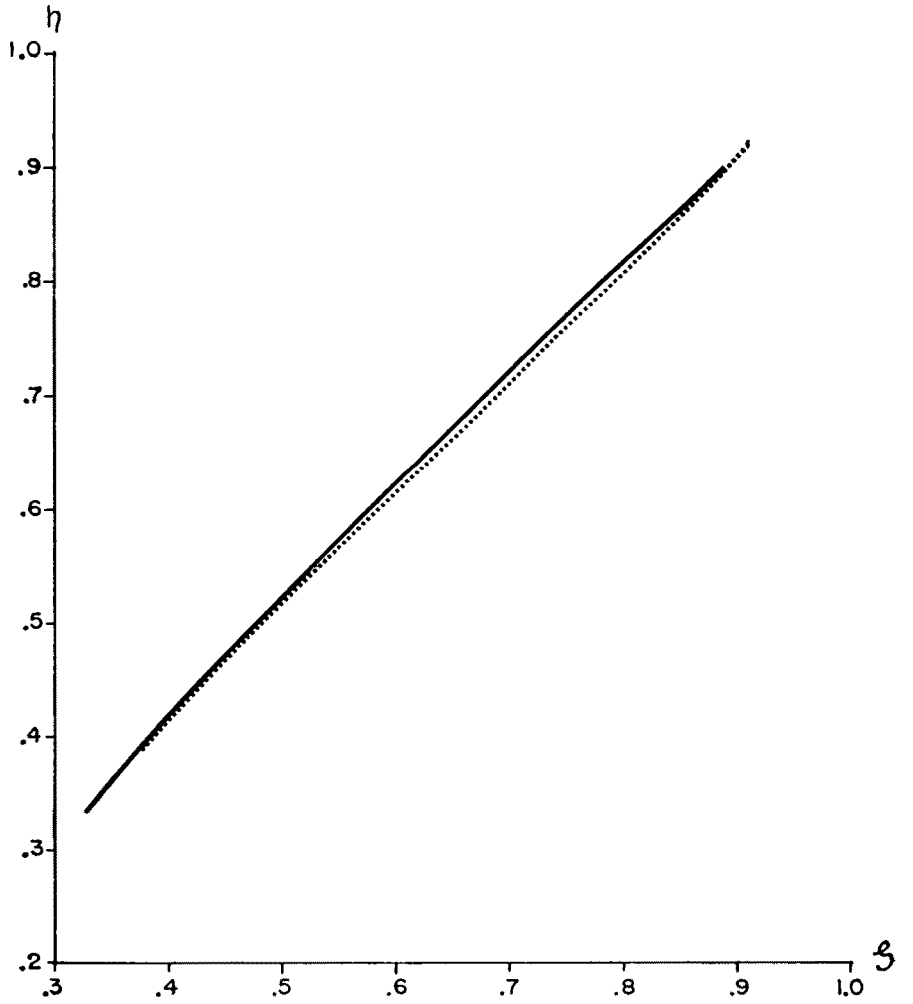

FIGURE 7
Functional relation between true scores on Tests P and Q as estimated from two rather different groups of examinees

Equating True Scores

Fig. 7 shows the estimated functional relation between true scores on two randomly parallel 60-item tests, P and Q, as determined from (11) within each of two somewhat different groups. The group statistics are as follows.

	$N = 2{,}000$	$N = 2{,}523$
Test P	$\bar{x} = 37.3;\ s_x^2 = 93.9$	$\bar{x} = 42.0;\ s_x^2 = 70.6$
Test Q	$\bar{x} = 38.5;\ s_x^2 = 92.6$	$\bar{x} = 42.7;\ s_x^2 = 67.1$

Since the test items were allocated to P and Q at random, the relation between their true scores should be almost linear, but not exactly so, as it would be if the tests were rigorously parallel. If P and Q measure the same traits regardless of group, then the functional relation obtained should be the same regardless of group except for sampling fluctuations. Fig. 7 is presented as a check on this point. Each curve runs from the first to the 99th percentile of the corresponding $g(\zeta)$. The agreement with expectation seems excellent.

Fig. 8 shows the functional relation between the true scores on Tests H and J as determined for the same two groups. The two curves would presumably be even more alike if the two tests did not measure slightly different types of vocabulary skill.

Simplified Model

The bivariate results described up to this point were obtained with k computed from (9). $\phi(x, y)$ was also obtained for seven of the same scatterplots, not involving Tests H and J, by setting $k = 0$. When (9) was used, none of the seven chi-squares were significant; with $k = 0$, all but one of the chi-squares were significant. This is the reason for distrusting the simplified model that sets $k = 0$.

Light is thrown on this failure by the following listing of predicted and observed product-moment correlations for the seven bivariate distributions studied.

Actual:	.860	.890	.886	.915	.798	.783	.783
$k \neq 0$:	.855	.885	.882	.910	.762	.776	.752
$k = 0$:	.833	.864	.865	.893	.747	.752	.717

It is clear that in every case, $k = 0$ gives much too low a correlation. This result can be predicted theoretically, since the use of $k = 0$ is the same as assuming that the correlation between parallel tests is equal to the Kuder-Richardson formula-21 reliability (see eq. 46), whereas the use of (9) replaces this with the formula-20 coefficient (see section 8). The differences just listed correspond very well to those commonly found between formulas 20 and 21.

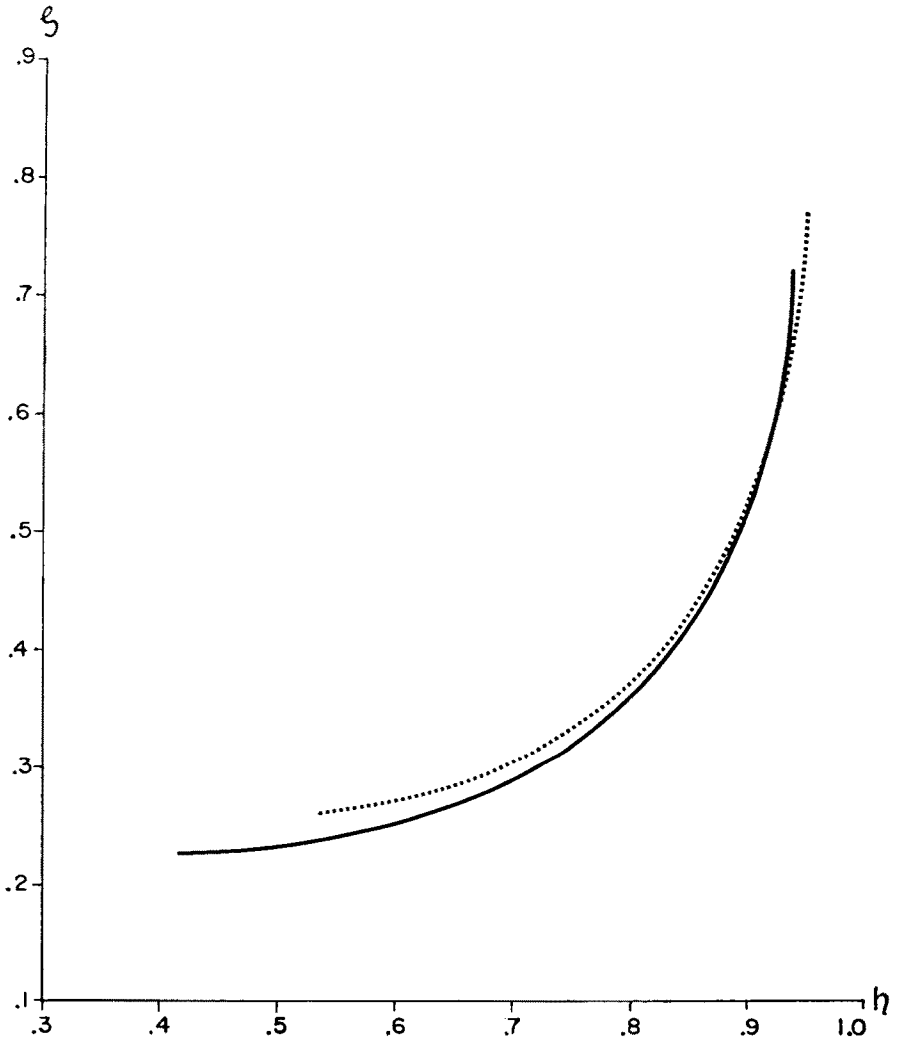

FIGURE 8
Functional relation between true scores on Tests H (ζ) and J (η) as estimated independently from two rather different groups of examinees

It will have been noted that computing k by (9) still gives a predicted correlation slightly too small. The difference is not enough to make each chi-square significant, but it is remarkably consistent from one scatterplot to the next. A result like this adds to rather than subtracts from the probable ultimate practical value of the model: it offers a chance to make the chi-square still better by finding an appropriate further adjustment to the

model—for example, a better method of estimating k, as discussed at the beginning of section 3.

The corresponding list of product-moment correlations for the four scatterplots between Tests H and J is given below for comparison.

Actual:	.642	.696	.607	.667
$k \neq 0$:	.664	.741	.644	.712
$k = 0$:	.656	(not computed)		

It is the comparison of this list with the preceding one that indicates that Test H and Test J measure slightly different vocabulary skills, since for this pair of tests the theoretical correlations are all too high, rather than too low, as found for the others.

The extension of the present model to deal with tests measuring different traits remains to be worked out.

Discussion

Occasionally an observed-score distribution is found such that routine application of the present model leads to an $a < 0$ or a $b > 1$. To date, this problem has been dealt with by setting $a = 0$ or $b = 1$, as appropriate, and fitting only three moments of the observed-score distribution instead of the usual four. So far, this method has been reasonably successful in the cases where it was required. If it should sometime fail, as seems likely, then some $g(\zeta)$ other than a beta distribution will be needed in order that the first four moments can all be fitted properly.

Other difficulties may arise if (i) the test is highly speeded, (ii) the test items are experimentally highly dependent, (iii) the test is "lumpy," consisting of two or more dissimilar clusters of items.

The univariate distributions successfully fitted to date have included some that are partly speeded, some that contain dependent items, and some that are lumpy. The bivariate distributions fitted so far have been chosen to avoid these difficulties. Further work will investigate the degree to which such difficulties can be tolerated. Modifications of the model to deal with such difficulties will also be investigated.

7. *Illustrative Application to a Selection Problem*

A group of 1000 examinees is to be divided into three subgroups by means of observed scores on a 25-item vocabulary test, x. The middle subgroup contains those examinees with scores $16 \leq x \leq 20$. What would be the distribution of scores for this middle subgroup on a parallel 25-item vocabulary test, y, if y could be immediately administered without practice effect?

As outlined in section 4, the answer is found by getting the predicted

$\hat{\phi}(x, y)$ and computing

$$\sum_{x=16}^{20} \hat{\phi}(x, y), \qquad y = 0, 1, \cdots, 25.$$

Illustrative results for one of the sets of $\hat{\phi}(x, y)$ discussed in section 6 are shown in Fig. 9. It is interesting to note that no trace of the peculiar truncated shape of the Test x distribution for the selected examinees is visually apparent in their predicted distribution on Test y.

In an application such as this, one can have confidence in the predicted distribution whenever one has confidence in the $\hat{\phi}(x, y)$, or has checked out the $\hat{\phi}(x, y)$ empirically as in the preceding section.

8. *Mathematical Derivations*

The Compound Binomial Distribution

Consider a group of examinees all having the same true score; suppose

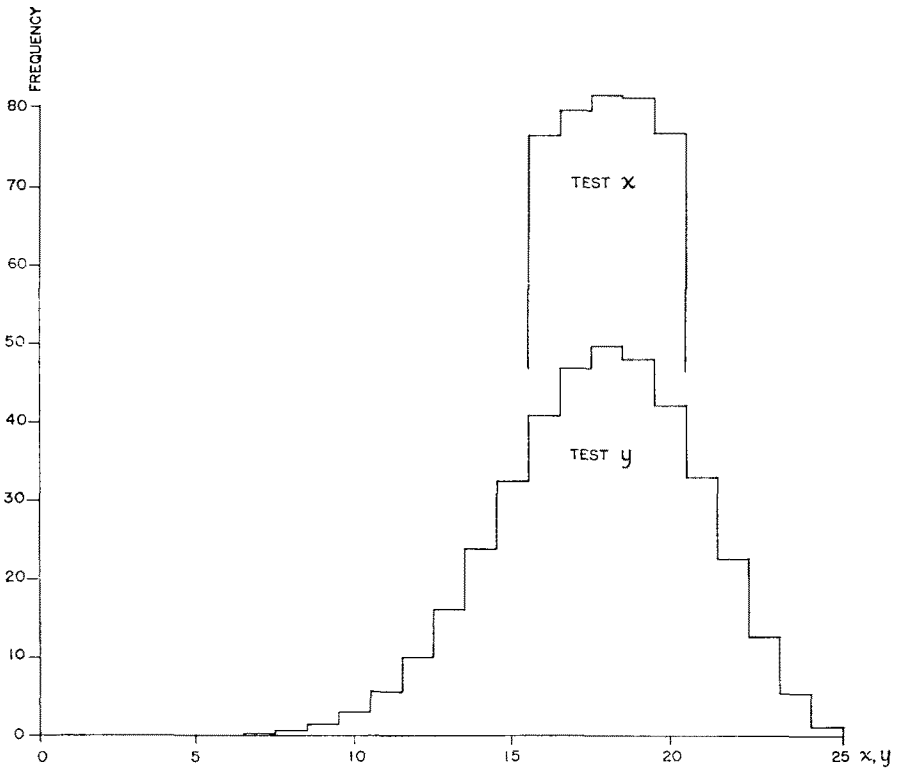

FIGURE 9
Observed distribution on Test x and estimated distribution on the parallel Test y after double truncation on Test x

the chance of success of any examinee on item i ($i = 1, 2, \cdots, n$) is p_i regardless of his performance on other items. Under these conditions, the examinee's test score, x, obtained by counting the number of his successes, is assumed to have a compound binomial distribution within the group specified (see, for example, [14], vol. 1, sect. 5.10). This compound binomial distribution has the n parameters $p_1, p_2, \cdots, p_n$.

Letting

$$p = \frac{1}{n}\sum_{i=1}^{n} p_i , \tag{15}$$

the compound binomial can be expanded in powers of $(p_1 - p)$, $(p_2 - p)$, $\cdots$, $(p_n - p)$, as outlined by Walsh and emended by the present writer [see 30, 5]. The compound binomial distribution can thus be written

$$\begin{aligned}\text{Prob}\,(X = x) = {} & p_n(x) + \frac{n}{2} V_2 C_2(x) \\ & + \frac{n}{3} V_3 C_3(x) + \left[\frac{n}{4} V_4 - \frac{n^2}{8} V_2^2\right] C_4(x) \\ & + \left[\frac{n}{5} V_5 - \frac{5n^2}{6} V_2 V_3\right] C_5(x) + \cdots , \qquad (x = 0, 1, \cdots, n),\end{aligned} \tag{16}$$

where

$$p_n(x) = \begin{cases} \binom{n}{x} p^x (1 - p)^{n-x}, & \text{for} \quad x = 0, 1, \cdots, n, \\ 0 \quad , & \text{otherwise,} \end{cases} \tag{17}$$

and

$$V_r = \frac{1}{n}\sum_{i=1}^{n} (p_i - p)^r , \qquad (r = 2, 3, \cdots, n). \tag{18}$$

Series (16) is a finite series; when all n terms are used, (16) is an exact identity. Since the higher-order terms become small, it is practical to truncate the series. A convenient result holds: no matter where (16) is truncated, the truncated series is a frequency distribution provided the parameters of the series are such that the truncated sum is nonnegative for $x = 0, 1, \cdots, n$.

Two-Term Approximation to the Compound Binomial

In the present work, only the first two terms of the series have been used:

$$\text{Prob}\,(X = x) = p_n(x) + \frac{n}{2} V_2 C_2(x), \qquad (x = 0, 1, \cdots, n). \tag{19}$$

Further simplification is necessary, however, since (19) depends on p and V_2, which in turn depend on the n unknown parameters $p_1, p_2, \cdots, p_n$.

For a given value of ζ, (19) can be thought of as involving just two parameters, p and V_2 ; however, these both vary with ζ. In this situation, as an approximation, we replace p and V_2 each by a convenient function of ζ.

An appropriate function for p arises rather convincingly from the following considerations. It can be readily shown (see eq. 33) that for (19) the expected value of x is np. Since $x/n - \zeta$ is thought of as an error of measurement, supposedly unbiassed, we want $E(x/n - \zeta \mid \zeta) = 0$, which is the same as $E(x/n \mid \zeta) = \zeta$ or

$$E(x \mid \zeta) = n\zeta. \tag{20}$$

It is seen that this last result will be achieved in (19) if we assume, as is done here, that p is the same as ζ:

$$p \equiv \zeta. \tag{21}$$

It is still necessary to find some approximation for V_2 . A crude approximation is used here: V_2 is assumed to be a quadratic function of ζ that vanishes (as it should) when $\zeta = 0$ or $\zeta = 1$. Specifically, it is assumed that

$$V_2 = \frac{2k}{n}\zeta(1 - \zeta), \tag{22}$$

where k is a parameter to be determined.

The conditional distribution of x for given ζ is thus for present purposes finally assumed to be as shown in (5).

From (5), (6), and (7) ,the two-term approximation to the compound binomial can be rewritten in various ways. With

$$\bar{\zeta} = 1 - \zeta, \tag{23}$$

we can write

$$P(x \mid \zeta) = p_n(x) + \frac{k}{n^{[2]}}\{-(x+1)(n-x-1)p_n(x+1) + 2x(n-x)p_n(x) - (x-1)(n-x+1)p_n(x-1)\}, \tag{24}$$

$$P(x \mid \zeta) = p_n(x)\left[1 + \frac{k}{\zeta\bar{\zeta}n^{[2]}}\{-\zeta^2(n-x)^{[2]} + 2\zeta\bar{\zeta}x(n-x) - \bar{\zeta}^2x^{[2]}\}\right], \tag{25}$$

$$P(x \mid \zeta) = p_n(x)\left[1 + \frac{k}{\zeta\bar{\zeta}n^{[2]}}\{-n^{[2]}\zeta^2 + 2x(n-1)\zeta - x^{[2]}\}\right], \tag{26}$$

where $x = 0, 1, \cdots, n$ and

$$u^{[v]} = u(u-1)\cdots(u-v+1). \tag{27}$$

The factorial moments of $p_n(x)$ are known ([14], vol. 1, eq. 5.11) to be

$$\mu_{[r]} = n^{[r]}\zeta^r, \qquad (r = 1, 2, \cdots, n). \tag{28}$$

If (26) is summed over all x, it is found with the aid of (28) that for all $0 < \zeta < 1$,

$$\sum_{x=0}^{n} P(x \mid \zeta) \equiv 1. \tag{29}$$

Next let us ask, for what values of x is $P(x \mid \zeta) \geq 0$. The term in square brackets in the third formulation of (26) is seen to be a quadratic in x, so oriented that its minimum in any interval on x must lie at one of the two end points of the interval. Thus $P(x \mid \zeta) \geq 0$ for all values of x in the interval bounded by the roots of this quadratic, which are

$$\begin{aligned} x &= \frac{1}{2} + (n-1)\zeta \pm \sqrt{\frac{1}{4} + \frac{1}{k}\zeta\bar{\zeta}(n-1)(n+k)} \\ &= \frac{1}{2} + n\zeta - \zeta \pm \sqrt{\frac{1}{k}\zeta\bar{\zeta}(n-1)(n+k)}\cdot\sqrt{1 + \frac{k}{4\zeta\bar{\zeta}(n-1)(n+k)}}. \end{aligned} \tag{30}$$

Expanding the final radical by Maclaurin series gives, after rearranging,

$$\begin{aligned} x = n\zeta \pm \sqrt{\frac{(n-1)(n+k)}{kn}}\sqrt{n\zeta\bar{\zeta}} + \frac{1}{2} - \zeta \\ \pm \frac{1}{8}\sqrt{\frac{k}{\zeta\bar{\zeta}(n-1)(n+k)}} \pm O(n^{-3/2}). \end{aligned} \tag{31}$$

The order of magnitude of the remainder was found to be $n^{-3/2}$ by using (9) with $\bar{x} = O(n)$, $\sigma_x^2 = O(n^2)$.

Now $n\zeta$ and $\sqrt{n\zeta\bar{\zeta}}$ are the mean and standard deviation of the binomial $p_n(x)$. Furthermore, it can be shown that whenever $.01 \leq \zeta \leq .99$, for all values of k and n so far encountered (see Table 1),

$$\left|\tfrac{1}{2} - \zeta \pm \sqrt{k}/8\sqrt{\zeta\bar{\zeta}(n-1)(n+k)}\right| < \tfrac{1}{2}.$$

Thus, for these values of ζ, k, and n, (31) states that $P(x \mid \zeta)$ is positive for all x in an interval which, except for a possible error of half a score unit plus a quantity of order $n^{-3/2}$, extends $\sqrt{(n-1)(n+k)/kn}$ standard deviations on each side of the mean of $p_n(x)$. Since $\sqrt{(n-1)(n+k)/kn} > 6$ for all tests so far considered, the value of $p_n(x)$ will be very small outside this interval. This suggests, in view of (26), that any negative "frequencies" in $P(x \mid \zeta)$ will be very small also, provided n is not too small, ζ is not too extreme, and k is not too large.

Values of $P(x \mid \zeta)$ were computed for a variety of n, k, and ζ in order to test the limits of this generalization. The results are summarized in Table 1, which suggests that $P(x \mid \zeta)$ is an extremely close approximation to a fre-

TABLE 1
Range Where $P(x|\zeta)$ Is Positive; Also Sum of All Negative "Frequencies"

n	k	$\sqrt{\frac{(n-1)(n+k)}{kn}}$	ζ	Values of x for which $P(x\|\zeta) \geq 0$	Sum of All Negative $P(x\|\zeta)$
50	2	5.0	.05	0 to 10	−.000004
60	4	4.0	.05	0 to 10	−.00007
50	4	3.6	.50	13 to 37	−.00003
25	2	3.6	.50	4 to 21	−.00002
50	4	3.6	.05	0 to 8	−.00022
25	2	3.6	.05	0 to 5	−.00037
50	4	3.6	.01	0 to 3	−.0008
25	2	3.6	.01	0 to 2	−.0013
50	8	2.7	.50	16 to 34	−.0021
50	8	2.7	.05	0 to 7	−.0026
50	8	2.7	.01	0 to 2	−.0043

quency distribution for most cases of interest. Difficulties might arise for $\zeta \leq .01$ or $\zeta \geq .99$ if n is too small or k too large. Even these difficulties can be ignored if the true-score distribution has little or no frequency for these extreme values of ζ.

Values of n and k for the 16 distributions reported in section 5 were as follows.

When $n = 25$: $k = 0.1, 0.1, 0.3, 0.4, 1.0, 1.1, 1.8, 2.1.$
When $n = 30$: $k = 1.6.$
When $n = 50$: $k = 1.0, 1.1, 3.9.$
When $n = 60$: $k = 3.6, 3.8, 3.9, 4.0.$

For these values, any negative frequencies in $P(x \mid \zeta)$ will be much too small to be of any concern, at least as long as $.01 \leq \zeta \leq .99$. None of the distributions studied has any appreciable frequency outside this range.

For convenience in writing, it is assumed that the values of n, k, and ζ dealt with are such that any negative values of $P(x \mid \zeta)$ are totally negligible; $P(x \mid \zeta)$ is referred to as a frequency distribution, without repeated qualifications.

The factorial moment of order r, denoted by $m_{[r]}(\zeta)$, can be found for $P(x \mid \zeta)$ by multiplying (26) by $x^{[r]}$, summing over all x, and using (28). After some algebra it is found that

$$(32) \qquad m_{[r]}(\zeta) = \sum_{x=0}^{n} x^{[r]} P(x \mid \zeta)$$
$$= n^{[r]}\zeta^r - kr^{[2]}(n-2)^{[r-2]}\zeta^{r-1}\bar{\zeta}, \qquad (r = 1, 2, \cdots, n),$$

where $(n-2)^{[r-2]} = 1$ if $r = 2$, and where the last term on the right vanishes for $r = 1$.

The mean and variance are found from (32) to be

$$E(x \mid \zeta) = n\zeta, \tag{33}$$

$$\text{var}\,(x \mid \zeta) = (n - 2k)\zeta\bar{\zeta}. \tag{34}$$

The Moments of the Distribution of True Scores

Multiply (1) by $x^{[r]}$ and sum over all x, placing the summation sign under the integral sign. If $h(x \mid \zeta) = P(x \mid \zeta)$, it will be seen that $M_{[r]}$, the rth factorial moment of $\phi(x)$, is

$$M_{[r]} = \int_0^1 g(\zeta) m_{[r]}(\zeta)\, d\zeta, \qquad (r = 1, 2, \cdots, n). \tag{35}$$

Substitute (32) in (35), writing μ'_r for the rth ordinary moment of $g(\zeta)$, to obtain

$$M_{[r]} = n^{[r]}\mu'_r - kr^{[2]}(n-2)^{[r-2]}(\mu'_{r-1} - \mu'_r), \qquad (r = 2, 3, \cdots, n). \tag{36}$$

This leads to a recurrence relationship, valid for any $g(\zeta)$ for which (1) holds and the moments exist:

$$\mu'_r = \frac{\dfrac{M_{[r]}}{(n-2)^{[r-2]}} + kr^{[2]}\mu'_{r-1}}{n^{[2]} + kr^{[2]}}, \qquad (r = 2, 3, \cdots, n); \tag{37}$$

also, from (35)

$$\mu'_1 = \frac{1}{n} M'_1, \tag{38}$$

where M'_1 is the mean of $\phi(x)$. Equations (37) and (38) enable us to obtain the moments of $g(\zeta)$ from the parameter k and the moments of $\phi(x)$. In particular, it is found that the true-score variance is

$$\mu_2 = \frac{\sigma_x^2 - (n - 2k)\bar{p}\bar{q}}{n^{[2]} + 2k}, \tag{39}$$

where, to conform to customary notation, σ_x^2 is the variance of $\phi(x)$, $\bar{p} = M'_1/n$ is the average item difficulty, and $\bar{q} = 1 - \bar{p}$.

Determining the Parameter k

The first product moment of observed score and true score is, from (1),

$$\mu'_{11} = \int_0^1 \zeta g(\zeta) \sum_{x=0}^{n} x h(x \mid \zeta)\, d\zeta. \tag{40}$$

By (40) and (21),

$$\mu'_{11} = n \int_0^1 \zeta^2 g(\zeta)\, d\zeta \tag{41}$$

$$= n\mu'_2 .$$

Thus the covariance between x and ζ is

$$\mu_{11} = n\mu'_2 - n\mu'^2_1 = n\mu_2 , \tag{42}$$

and their correlation is

$$r_{x\zeta} = \frac{n\sqrt{\mu_2}}{\sigma_x}. \tag{43}$$

Equation (43) is a familiar formula for the index of reliability.

The usual Kuder-Richardson formula-20 reliability statistic (r_{20}) will be used as an estimate of $r^2_{x\zeta}$. In the form given by Tucker [28], this estimate is

$$r^2_{x\zeta} \doteq r_{20} = \frac{n}{n-1}\left[1 - \frac{n(\bar{p}\bar{q} - s_p^2)}{s_x^2}\right] \tag{44}$$

(see (9) for an explanation of the notation). Eliminate $r_{x\zeta}$ from (43) and (44) and replace σ_x^2 by s_x^2, obtaining approximately

$$n^2\mu_2 \doteq \frac{n}{n-1}\,[s_x^2 - n(\bar{p}\bar{q} - s_p^2)]. \tag{45}$$

Eliminate μ_2 from (39) and (45) and solve for k, obtaining the approximation

$$k \doteq \frac{n^2(n-1)s_p^2}{2[n^2\bar{p}\bar{q} - s_x^2 - ns_p^2]}. \tag{46}$$

Since $\bar{p} = \bar{x}/n$, this result is the same as the estimate of k given by (9).

For the usual range of $\bar{p}$, an approximate value of k is $k \doteq 2ns_p^2$. The test reliability can be computed from k by the formula

$$r_{20} = \frac{n}{n - 1 + \dfrac{2k}{n}}\left[1 - \frac{\bar{p}\bar{q}(n-2k)}{s_x^2}\right]. \tag{47}$$

If $k = 0$, (47) becomes a Kuder-Richardson formula-21 reliability.

Equation (34) suggests that when the model holds, k will tend to remain roughly the same from group to group. This conclusion follows if one accepts the usual assumption that var $(x \mid \zeta)$ remains constant from group to group provided the groups are not so dissimilar that the meaning of ζ changes from one to another.

Once k is computed from (9), then the first four moments of $g(\zeta)$ can be estimated by means of (37) and (38), using the moments of the observed $f(x)$ as estimates of the moments of $\phi(x)$. The parameters a, b, α, β of the

four-parameter beta distribution can then be estimated from the first four estimated moments of $g(\zeta)$, using standard formulas [e.g., 6].

Estimating $\phi(x)$ When $k = 0$

Explicit formulas are available for the $\phi(x)$ given in (8). It is simplest to discuss first the case where $k = 0$, denoting the result by $\phi_0(x)$. In this case, (8) can be written in a form involving a bivariate hypergeometric function ([7], eq. 5.8.2(5)):

$$(48)\quad \phi_0(x) = \binom{n}{x} a^x (1-a)^{n-x} F_1\left(\alpha, -x, -n+x, \alpha+\beta; -\frac{A}{a}, \frac{A}{1-a}\right)$$

$$= \binom{n}{x} a^x (1-a)^{n-x} \sum_{i=0}^{n} \sum_{j=0}^{n} \frac{(\alpha)_{i+j}(-x)_i(-n+x)_j}{(\alpha+\beta)_{i+j} i!j!} \cdot \left(-\frac{A}{a}\right)^i \left(\frac{A}{1-a}\right)^j, \qquad (x = 0, 1, \cdots, n).$$

By using the relationships ([7], eq. 5.11(5))

$$(49)\quad F_1(\alpha, \beta, \beta', \gamma; z, y) = (1-z)^{-\beta}(1-y)^{\gamma-\alpha-\beta'} \cdot F_1\left(\gamma-\alpha, \beta, \gamma-\beta-\beta', \gamma; \frac{z-y}{z-1}, y\right)$$

and ([1], p. 24, eq. (29′), here modified to correct minor errors)

$$(50)\quad F_1(\alpha, \beta, \beta', \gamma; z, y) = (1-z)^{-\beta} F_3\left(\gamma-\alpha, \alpha, \beta, \beta', \gamma; \frac{z}{z-1}, y\right),$$

where

$$(51)\quad F_3(\alpha, \alpha', \beta, \beta', \gamma; z, y) = \sum_{i=0}^{\infty} \frac{(\alpha)_i (\beta)_i}{(\gamma)_i i!} F(\alpha', \beta'; \gamma+i; y) z^i$$

and $F(\alpha, \beta; \gamma: y)$ is the usual one-variable hypergeometric function, it is found that

$$(52)\quad \phi_0(x) = \sum_{i=0}^{n} \sum_{j=0}^{n} \left[\frac{n! a^{x-i}}{(n-x)!i!(x-i)!(1-a)^x}\right] \cdot \left[\frac{(\alpha)_i A^j (n-i)^{[j]} (\beta)_j (1-b)^{n-i-j}}{(\alpha+\beta)_{i+j} j!}\right][A^j], \qquad (x = 0, 1, \cdots, n).$$

This expression has the computational advantage over (48) in that all terms under the summation signs are nonnegative. This avoids the serious loss of significant digits that occurs when terms of opposite sign in (48) are added together. The brackets are introduced into (52) to indicate how the $\phi_0(x)$, $x = 0, 1, \cdots, n$, may be computed as the elements of a matrix product. The typical element of the first matrix is shown in the first bracket, x being the row subscript and i the column subscript; this matrix has only

zero elements above the main diagonal. The second matrix has row subscript i and column subscript j; it has zeros below the secondary diagonal. The third matrix is a column vector with subscript j. Within any one matrix, each element after the first is conveniently computed from an adjacent element.

Estimating $\phi(x)$ When $k \neq 0$

If (24) is substituted for $C_2(x)$ in (8), it is seen that

$$\phi(x) = \phi_0(x) - \frac{k}{n^{[2]}} \Delta^2_{x-1}, \qquad (x = 0, 1, \cdots, n), \tag{53}$$

where

$$\Delta^2_{x-1} = \Psi(x-1) - 2\Psi(x) + \Psi(x+1) \tag{54}$$

and

$$\Psi(x) = x(n-x)\phi_0(x), \tag{55}$$

with $\phi_0(x) = 0$ for $x < 0$ and for $x > n$.

Estimating $\phi(x, y)$

The first step in evaluating (13) is to determine $\psi(\zeta)$ from (11) for numerous values of ζ. This is currently done on the electronic computer using a combination of numerical integration and inverse interpolation. Then (13) is evaluated with $k = 0$, obtaining $\phi_0(x, y)$ for each x and y by numerical integration.

The value of $\phi(x, y)$ for $k \neq 0$ is then found from (56), which was derived by the same approach used to obtain (53):

$$\begin{aligned}\phi(x, y) = \phi_0(x, y) &+ \frac{k_x}{n^{[2]}} \{-\psi_v(x+1) + 2\psi_v(x) - \psi_v(x-1)\} \\ &+ \frac{k_y}{m^{[2]}} \{-\psi_x(y+1) + 2\psi_x(y) - \psi_x(y-1)\} \\ &+ \frac{k_x k_y}{n^{[2]} m^{[2]}} \{\psi(x+1, y+1) - 2\psi(x+1, y) + \psi(x+1, y-1) \\ &- 2\psi(x, y+1) + 4\psi(x, y) - 2\psi(x, y-1) + \psi(x-1, y+1) \\ &- 2\psi(x-1, y) + \psi(x-1, y-1)\},\end{aligned} \tag{56}$$

where

$$\begin{aligned}\psi_v(x) &= x(n-x)\phi_0(x, y), \\ \psi_x(y) &= y(m-y)\phi_0(x, y), \\ \psi(x, y) &= x(n-x)y(m-y)\phi_0(x, y),\end{aligned} \tag{57}$$

k_x and k_y are the values of k for tests x and y, and m is the number of items in test y. Also, $\phi_0(x, y) = 0$ if either $x < 0$, $y < 0$, $x > n$, or $y > m$.

REFERENCES

[1] Appell, P. and Kampé de Fériet, J. Fonctions hypergeometriques et hyperspheriques. Polynomes d' Hermite. Gauthier-Villars, 1926.

[2] Buckingham, R. A. *Numerical methods.* London: Pitman, 1957.

[3] Bückner, H. F. Numerical methods for integral equations. In J. Todd (Ed.), *A survey of numerical analysis.* New York: McGraw-Hill, 1962.

[4] Cochran, W. G. Some methods for strengthening the common χ^2 tests. *Biometrics,* 1954, **10,** 417–451.

[5] Eisenberg, H. B., Geoghagen, R. R. M., and Walsh, J. E. A general probability model for binomial events with application to surgical mortality. *Biometrics,* 1963, **19,** 152–157.

[6] Elderton, Sir W. P. *Frequency curves and correlation.* (3rd ed.). Cambridge, England: Cambridge Univ. Press, 1938.

[7] Erdélyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G. *Higher transcendental functions.* Vol. 1. New York: McGraw-Hill, 1953.

[8] Fox, L. and Goodwin, E. T. The numerical solution of non-singular linear equations. *Phil. Trans. roy. Soc. London (Series A),* 1952–53, **245,** 501–535.

[9] Gulliksen, H. O. *Theory of mental tests.* New York: Wiley, 1950.

[10] Hirschman, I. I. and Widder, D. V. *The convolution transform.* Princeton: Princeton Univ. Press, 1955.

[11] Jeffries, J. and Orrall, F. The numerical solution of Fredholm integral equations of the first kind. (Unclassified report reproduced by the Armed Services Technical Information Agency.) Arlington, Va.: Arlington Hall Station. November 1960. AD 250–862.

[12] Kantorovich, L. W. and Krylov, V. I. *Approximate methods of higher analysis.* New York: Interscience, 1958.

[13] Keats, J. A. and Lord, F. M. A theoretical distribution for mental test scores. *Psychometrika,* 1962, **27,** 59–72.

[14] Kendall, M. G. and Stuart, A. *The advanced theory of statistics.* New York: Hafner, 1958.

[15] Kunz, K. S. *Numerical analysis.* New York: McGraw-Hill, 1957.

[16] Lord, F. M. A theory of test scores. *Psychometric Monogr.,* No. 7. Richmond, Va.: William Byrd Press, 1952.

[17] Lord, F. M. An approach to mental test theory. *Psychometrika,* 1959, **24,** 283–302.

[18] Lord, F. M. Use of true-score theory to predict moments of univariate and bivariate observed-score distributions. *Psychometrika,* 1960, **25,** 325–342.

[19] Lord, F. M. Estimating true measurements from fallible measurements (Binomial Case) —expansion in a series of beta distributions. Res. Bull. 62–23. Princeton, N. J.: Educ. Test. Serv., 1962.

[20] Lord, F. M. True-score theory—The four parameter beta model with binomial errors. Res. Bull. 64–6. Princeton, N. J.: Educ. Test. Serv., 1964.

[21] Medgyessy, P. *Decomposition of superpositions of distribution functions.* Budapest: Publishing House of the Hungarian Academy of Sciences, 1961.

[22] Phillips, D. L. A technique for the numerical solution of certain integral equations of the first kind. *J. Ass. comput. Machinery,* 1962, **9,** 84–97.

[23] Pollard, H. Distribution functions containing a Gaussian factor. *Proc. Amer. math. Soc.,* 1953, **4,** 578–582.

[24] Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen and Lydiche (M. Simmelkiaer), 1960.

[25] Teicher, H. On the mixture of distributions. *Ann. math. Statist.*, 1960, **31**, 55–73.
[26] Tricomi, F. G. *Integral equations.* New York: Interscience, 1957.
[27] Trumpler, R. J. and Weaver, H. F. *Statistical astronomy.* Berkeley: Univ. California Press, 1953.
[28] Tucker, L. R. A note on the estimation of test reliability by the Kuder-Richardson formula (20). *Psychometrika*, 1949, **14**, 117–120.
[29] Twomey, S. On the numerical solution of Fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature. *J. Ass. comput. Machinery*, 1963, **10**, 97–101.
[30] Walsh, J. E. Corrections to two papers concerned with binomial events. *Sankhyā*, Ser. A., 1963, **25**, 427.
[31] Walther, A. and Dejon, B. General report on the numerical treatment of integral and integro-differential equations. In *Symposium on the Numerical Treatment of Ordinary Differential Equations, Integral and Integro-Differential Equations—Proceedings of the Rowe Symposium (20–24 September 1960).* Basel/Stuttgart: Birkhauser Verlag, 1960. Pp. 645–671.
[32] Young, A. Approximate product-integration. *Proc. roy. Soc.*, A, 1954, **224**, 552–561.
[33] Young, A. The application of product-integration to the numerical solution of integral equations. *Proc. roy. Soc.*, A, 1954, **224**, 561–573.

Manuscript received 6/20/64
Revised manuscript received 8/7/64