RB-67-37

ESTIMATING TRUE-SCORE DISTRIBUTIONS IN PSYCHOLOGICAL TESTING (AN EMPIRICAL BAYES ESTIMATION PROBLEM)

Frederic M. Lord

Office of Naval Research Contract Nonr-2752(00)
Project Designation NR 151-201
Frederic M. Lord, Principal Investigator

Educational Testing Service
Princeton, New Jersey

November 1967

RB-67-37

ESTIMATING TRUE-SCORE DISTRIBUTIONS IN PSYCHOLOGICAL TESTING (AN EMPIRICAL BAYES ESTIMATION PROBLEM)

Frederic M. Lord

Office of Naval Research Contract Nonr-2752(00)
Project Designation NR 151-201
Frederic M. Lord, Principal Investigator

Educational Testing Service
Princeton, New Jersey

November 1967

ESTIMATING TRUE-SCORE DISTRIBUTIONS IN PSYCHOLOGICAL TESTING (AN EMPIRICAL BAYES ESTIMATION PROBLEM)

Abstract

The following problem is considered: Given that the frequency distribution of the errors of measurement is known, determine or estimate the distribution of true scores from the distribution of observed scores for a group of examinees. Typically this problem does not have a unique solution. However, if the true-score distribution is "smooth", then any two smooth solutions to the problem will differ little from each other. Methods for finding smooth solutions are developed a) for a population and b) for a sample of examinees. The results of a number of tryouts on actual test data are summarized.

ESTIMATING TRUE-SCORE DISTRIBUTIONS IN PSYCHOLOGICAL TESTING (AN EMPIRICAL BAYES ESTIMATION PROBLEM)*

When a large group of individuals has been tested, the examiner usually finds the frequency distribution of observed test scores to be of some interest. However, he would usually prefer to look at the frequency distribution of true scores, if this were possible. Is the true-score distribution bimodal? For a multiple-choice test, do some individuals have true scores below the "chance" level (the score that would be expected if they responded entirely at random)?

Although an estimated true-score distribution is of interest for itself, it is more often of practical value as an intermediate step in the prediction of more tangible results. As pointed out by Lord (1965), the estimated true-score distribution "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).

---

*The writer wishes to thank Diana Lees and Virginia Lennon, who wrote the computer programs, carried out some of the mathematical derivations, and helped with other important aspects of the work. This work was supported in part by contract Nonr-2752(00) between the Office of Naval Research and Educational Testing Service. Reproduction, translation, use and disposal in whole or in part by or for the United States Government is permitted.

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."

An additional use, of some interest, is

9. To estimate the item-true score regression for particular items, without strong prior assumption as to its mathematical form.

Practical applications of true-score theory will not be discussed further here. The present article defines the problem (section 1), outlines some of the obstacles to a satisfactory solution (sections 2, 5), suggests some solutions to the mathematical problem (sections 3, 4) and to the related statistical problem (section 6). Some empirical checks of actual results are described and discussed (sections 8-13). Certain of the mathematical details are spelled out in the appendix.

1. The Basic Mathematical Model

Let $x = 0,1,\ldots,n$ be the number of right answers given by an examinee on an n-item test; let ξ, $0 \leq \xi \leq n$, be the <u>true score</u> of the examinee. The mathematical formulation will be in terms of $\zeta \equiv \xi/n$, which also will be called the <u>true score</u>. [The identity sign will be used to denote a definition, as well as with its usual meanings.]

Let $\phi(x)$ denote the proportion of the population of examinees with observed score x, let $h(x|\zeta)$ denote the corresponding conditional proportions for fixed true score, and let $g(\zeta)$ be the noncumulative frequency distribution of true scores. Ordinarily ζ is a continuous variable, covering the range $0 \leq a \leq \zeta \leq b \leq 1$ (an $a > 0$ and a $b < 1$, while not mathematically essential, are sometimes helpful in practical applications). It follows that

$$\phi(x) = \int_a^b g(\zeta)\, h(x|\zeta)\, d\zeta \tag{1}$$

for $x = 0,1,\ldots,n$. The basic problem is: Given some $h(x|\zeta)$ and an observed-score distribution $\phi(x)$, infer from (1) the true-score distribution $g(\zeta)$.

Most of the theoretical results to be obtained will be written down without specifying the form of $h(x|\zeta)$. To obtain practical results, however, it is necessary that the form of $h(x|\zeta)$ be known. As in Lord (1965), it is assumed that $h(x|\zeta)$ is a compound binomial distribution. In actuality, applied results are obtained by using a four-term Taylor series approximation to the compound binomial (Lord, 1965, eq. 16). The details necessary for computing this approximation, given by Lord and Lees (1967a), will not be considered further here.

Lord (1965) assumed that $g(\zeta)$ was a four-parameter beta distribution.

The range of this distribution is given by two parameters, a and b, having the same meaning as in (1). It now appears from studies of widely varied test administrations (Lord & Lees, 1967a) that this assumption works well when the estimates $\hat{a}$ and $\hat{b}$ obtained from the observed-score distribution fall in the permissible range $0 \leq \hat{a} < \hat{b} \leq 1$. When $\hat{a} < 0$ or $\hat{b} > 1$, however, the obvious reestimation procedure under the requirement that $\hat{a} = 0$ or $\hat{b} = 1$ frequently does not yield good results. This is hardly surprising since in such cases one is in effect fitting only a three- rather than a four-parameter distribution.

It seems preferable to start without assuming a specified mathematical form for $g(\zeta)$. Let us see what can be done without any such assumption.

2. Multiplicity of Solutions*

If x were a continuous variable and if equation (1) held for all values of x in some interval, then (1) would be a Fredholm integral equation of the first kind (e.g., see Tricomi, 1957). Any function $g(\zeta)$ satisfying (1) is called a solution to the integral equation. In the actual case, x is limited to the integers $0,1,\ldots,n$. Let $\tilde{x}$ be a continuous variable and let $\Psi(\tilde{x})$ be any continuous function of $\tilde{x}$ in the interval $0 \leq \tilde{x} \leq n$ such that $\Psi(\tilde{x}) = \phi(x)$ when $\tilde{x} = x = 0,1,\ldots,n$.

---

*Section 2 and major portions of sections 3 and 5 are abstracted, with minor revisions, from Lord and Lees (1967b).

Then any solution to

$$\Psi(\tilde{x}) = \int_a^b g(\zeta)\, h(\tilde{x}|\zeta)\, d\zeta$$

is automatically a solution to (1). If there is any solution to (1), there will in general be an infinite number of solutions when x is integer-valued.

If $h(x|\zeta)$ is binomial, the first n moments of $g(\zeta)$ are given (Skellam, 1948) by

$$\mu'_r = \frac{M_{[r]}}{n^{[r]}} \qquad (r = 1,2,\ldots,n) \quad , \tag{2}$$

where $M_{[r]}$ is the r-th factorial moment of $\phi(x)$ and

$$n^{[r]} \equiv n(n-1)\ldots(n-r+1) \quad . \tag{3}$$

A similar but more complicated result holds for the four-term series approximation to the compound binomial used here (see Lord & Lees, 1967a, eq. 41). Thus the first n moments of the observed-score distribution are determined by the first n moments of the true-score distribution. Since the frequency distribution of a bounded integer-valued variable is determined by its moments (see Riordan, 1958, ch. 2, eq. 32), it follows that <u>any true-score distribution with the proper moments up through order n will be a solution to the integral equation (1)</u>. The same statement holds if $h(x|\zeta)$ is compound binomial. Thus (again), even given an infinite number of observations, it is impossible by means of (1) to determine the true-score distribution from the observed distribution of number-right scores.

Is there some further criterion that can be used to determine the true-score distribution? In almost any practical situation we would expect the true-score distribution to be "smooth" in some vaguely defined sense, best evaluated by visual inspection. Thus we may require that the solution to (1) shall not be obviously irregular. Among other things, we may require that it be unimodal. These restrictions still will not provide a unique method of determining the true-score distribution.

However, consider any two "smooth" solutions, $g_1(\zeta)$ and $g_2(\zeta)$, to equation (1). Since these two distributions have the same moments up through order n, they also have the same best-fitting polynomial of degree n in the least-squares sense (Kendall & Stuart, 1958, sect. 3.34). Denote this polynomial by $P(\zeta)$. Typically, $n \geq 15$. Now if a distribution is "smooth" in the ways ordinarily expected, without peculiar irregularities, it should be possible to fit it very closely by a polynomial of degree $n \geq 15$. We would expect to find that $\epsilon_1 \equiv \underset{\zeta}{\text{Max}}\, g_1(\zeta) - P(\zeta)$ and $\epsilon_2 \equiv \underset{\zeta}{\text{Max}}\, g_2(\zeta) - P(\zeta)$ are both very small quantities. This seems a natural part of what is meant by "smoothness". Consequently, $g_1(\zeta)$ and $g_2(\zeta)$ can differ at most by $|\epsilon_1| + |\epsilon_2|$, a small quantity.

In summary, we can not hope to determine the true-score distribution uniquely. Given enough observations, however, we can reasonably hope that for the values of n encountered in practice, any acceptable solution will differ from any other acceptable solution by an amount negligible for most practical purposes.

3. Solving the Integral Equation

Equation (1) is really $n + 1$ equations--one for each integer $0,1,\ldots,n$. Choose any function $t(u,\zeta)$ such that i) the integrals

$$m_{ux} = \int_a^b t(u,\zeta)\, h(x|\zeta)\, d\zeta \qquad (u,x = 0,1,\ldots,n) \tag{4}$$

exist; ii) the inverse of the matrix $\|m_{ux}\|$ exists. It will be seen in later sections that functions $t(u,\zeta)$ satisfying these two conditions can be found, at least for the case where $h(x|\zeta)$ is binomial or compound binomial.

Let m^{xu} denote a typical element of the inverse of $\|m_{ux}\|$. The function

$$g(\zeta) = \sum_{u=0}^{n} w_u t(u,\zeta) \tag{5}$$

will satisfy equation (1) provided the weights w_u are obtained from

$$w_u = \sum_{v=0}^{n} m^{vu}\phi(v) \qquad (u = 0,1,\ldots,n) . \tag{6}$$

To prove this, substitute (5) and (6) into (1), obtaining after some rearrangement

$$\phi(x) = \sum_{v=0}^{n} \phi(v) \sum_{u=0}^{n} m^{vu} \int_a^b t(u,\zeta)\, h(x|\zeta)\, d\zeta .$$

Use (4) and let

$$\delta_{vx} \equiv \begin{cases} 1 & \text{if } v = x \\ 0 & \text{if } v \neq x \end{cases} , \tag{7}$$

finding that

$$\begin{aligned} \phi(x) &= \sum_{v=0}^{n} \phi(v) \sum_{u=0}^{n} m^{vu} m_{ux} \\ &= \sum_{v=0}^{n} \delta_{vx} \phi(v) \\ &= \phi(x) \qquad (x = 0,1,\ldots,n) , \end{aligned}$$

which completes the proof.

Thus for any given observed-score distribution $\phi(x)$ $(x = 0,1,\ldots,n)$, for any conditional distribution $h(x|\zeta)$, and for any chosen function $t(u,\zeta)$ satisfying the stated conditions, one can determine $\|m_{ux}\|$ from (4), invert this matrix, determine the weights w_u from (6), and use (5) to write down a $g(\zeta)$ satisfying equation (1). In the special applications to be treated here, it will be seen that $t(u,\zeta)$ and the solution $g(\zeta)$ are polynomials in ζ.

If the $g(\zeta)$ so determined is nonnegative in the range $a \leq \zeta \leq b$, then it is automatically a frequency distribution satisfying the condition

$$\int_a^b g(\zeta)\, d\zeta = 1 . \tag{8}$$

We can prove this by summing (1) on x, obtaining

$$\sum_{x=0}^{n} \phi(x) = 1 = \int_a^b g(\zeta) \sum_{x=0}^{n} h(x|\zeta)\, d\zeta = \int_a^b g(\zeta)\, d\zeta .$$

4. The Smoothest Solution

If one believes that the true-score distribution is smooth, a good approximation to it might be found by choosing, from among the infinite number of solutions to (1), the solution that is smoothest in some sense. First some measure of "smoothness" is necessary. There does not seem to be any uniquely good way to define smoothness. A cover-all measure of the unsmoothness of a function $g(\zeta)$ is

$$\bar{S} \equiv \sum_{r=0}^{R} \int_{a}^{b} w_r(\zeta) \, [g^{(r)}(\zeta)]^2 \, d\zeta \tag{9}$$

where $g^{(r)}(\zeta)$ is the r-th derivative of $g(\zeta)$ and where the $w_r(\zeta) \geq 0$ are weighting functions at the disposal of the statistician.

After choosing some $w_r(\zeta)$, $r = 0,1,\ldots,R$, we then try to find the $g(\zeta)$ that will minimize $\bar{S}$ subject to the restriction that $g(\zeta)$ satisfies (1) for $x = 0,1,\ldots,n$. This is a problem in the calculus of variations (e.g., see Pars, 1962), which, by Euler's rule, is equivalent to the problem of choosing $g(\zeta)$ so as to minimize

$$\int_{a}^{b} \left(\sum_{r=0}^{R} w_r(\zeta) \, [g^{(r)}(\zeta)]^2 - 2 \sum_{x=0}^{n} \lambda_x \, g(\zeta) \, h(x|\zeta) \right) d\zeta \quad , \tag{10}$$

where the λ_x (like Lagrange multipliers) are constants to be determined.

Consider the simplest case where $R = 0$, in which case

$$\bar{S} = \int_a^b w(\zeta)\,[g(\zeta)]^2\,d\zeta \quad . \tag{11}$$

Define $\gamma(\zeta) \equiv 1/w(\zeta)$. Since $\gamma(\zeta)$ is at our disposal, we may without loss of generality require that $\int_a^b \gamma(\zeta)\,d\zeta = 1$, in which case $\gamma(\zeta)$ is automatically a frequency distribution on the interval (a,b) . Since $\int_a^b g(\zeta)\,d\zeta = 1$ also, it is easily found (after expansion) that

$$\int_a^b \frac{[g(\zeta) - \gamma(\zeta)]^2}{\gamma(\zeta)}\,d\zeta \tag{12}$$

differs from $\bar{S}$ by a constant. This last expression is a familiar distance measure, being the analog for continuous frequency distributions of a chi square between $g(\zeta)$ and $\gamma(\zeta)$.

The foregoing result suggests that we should choose $\gamma(\zeta)$ to be some smooth first approximation to the true $g(\zeta)$. The procedure of minimizing $\bar{S}$ will then make the estimated $g(\zeta)$ as much like $\gamma(\zeta)$ as possible, in the metric defined by (12), while still satisfying (1) for all x . In particular, when $\gamma(\zeta)$ is near zero, the presence of $\gamma(\zeta)$ in the denominator of (12) will force the difference $g(\zeta) - \gamma(\zeta)$ to be very small.

In practical work, we shall frequently take $\gamma(\zeta)$ to be a rectangular distribution, that is, $\gamma(\zeta) = \text{constant}$. If the resulting estimated $g(\zeta)$ vanishes at $\zeta = 0$ and at $\zeta = 1$, then the estimate obtained is not changed drastically by using the triangular distribution $\gamma(\zeta) \propto \zeta$ or

$\gamma(\zeta) \propto 1 - \zeta$ instead of $\gamma(\zeta) \propto 1$. The same is true of the parabola $\gamma(\zeta) \propto \zeta(1 - \zeta)$. Bell-shaped distributions such as $\gamma(\zeta) \propto \zeta^2(1 - \zeta)^2$ will sometimes be used in work with fallible data to prevent the appearance of implausible bumps in the tails of estimated true-score distributions.

When $R = 0$, (10) becomes

$$(13) \qquad \int_a^b \left\{ w(\zeta)\,[g(\zeta)]^2 - 2 \sum_{x=0}^{n} \lambda_x\, g(\zeta)\, h(x|\zeta) \right\} d\zeta \quad .$$

A necessary condition for finding $g(\zeta)$ to minimize (13), thus minimizing (9) subject to (1), is that $g(\zeta)$ satisfy the Euler equation

$$(14) \qquad g(\zeta) = \gamma(\zeta) \sum_{x=0}^{n} \lambda_x\, h(x|\zeta)$$

(obtained in this simple case by treating g in (13) as an independent variable, differentiating the integrand with respect to g, and setting this derivative equal to zero). The $n + 1$ values of λ_x are to be determined so that (1) is satisfied for $x = 0,1,\ldots,n$. Under general regularity conditions it can be shown further (Pars, 1962, pp. 103-104) that (14) is sufficient for a minimum.

Given $\gamma(\zeta)$ and $h(x|\zeta)$, the $n + 1$ values of λ_x needed for (14) can be determined from the $n + 1$ values of $\phi(x)$, as follows. Replace x in (14) by X and substitute (14) into (1) to obtain

$$(15) \qquad \phi(x) = \sum_{X=0}^{n} \lambda_X m_{xX} \qquad (x = 0,1,\ldots,n)$$

where

$$(16) \qquad m_{xX} = \int_a^b \gamma(\zeta)\, h(x|\zeta)\, h(X|\zeta)\, d\zeta \qquad (x,X = 0,1,\ldots,n) \quad .$$

If the inverse of the matrix $\|m_{xX}\|$ exists, then the simultaneous linear equations of (15) can be solved uniquely for the λ_X :

$$(17) \qquad \lambda_X = \sum_{x=0}^{n} m^{Xx}\, \phi(x) \qquad (X = 0,1,\ldots,n) \quad ,$$

where m^{Xx} is the general element of the inverse of $\|m_{xX}\|$.

In order for the $g(\zeta)$ given by (14) and (17) to be useful here, it is of course necessary that it be nonnegative in the range $a \leq \zeta \leq b$. This requirement could be imposed as part of the calculus of variations problem (see Kenneth & Taylor, 1966, and Leitmann, 1962). Further discussion of this requirement will be deferred to later sections.

The restriction that $R = 0$ in (9) is clearly an oversimplification made to keep the analysis tractable. It is possible to proceed with $R \neq 0$, but no empirical work has been done for this more complicated case.

The reader should clearly understand that the problem equations (14) and (17) purport to solve is not ordinarily encountered in practical work. In practice, we never know the values of $\phi(x)$ required in (17); we only have sample frequencies that approximate $\phi(x)$. The problem where the population frequencies $\phi(x)$ are assumed known will be referred to as

the mathematical problem; the practical problem where only the sample frequencies, to be denoted by $f(x)$, are known will be referred to as the statistical problem. It will be seen that the obvious devices ordinarily used in statistical inference do not usually lead to an acceptable solution to the statistical problem.

5. Statistical Estimation Problems

In practical applications, the problem is to estimate $g(\zeta)$ in (1), given the sample frequency distribution $f(x)$. As before, the model assumes that $h(x|\zeta)$ is known--in the present application, that it is a certain kind of compound binomial distribution.

This kind of estimation problem is known as an empirical Bayes problem. The true-score distribution $g(\zeta)$ is the prior distribution. The problem of estimating the prior distribution is treated mathematically by Robbins (1964). Maritz (1966) uses the device of assuming $g(\zeta)$ to be a discrete distribution with ζ taking only a limited number of values. Here we prefer to try to make $g(\zeta)$ as smooth as possible, consistent with the observed data.

It is not uncommon to solve certain estimation problems first in terms of population parameters, after which substitution of sample statistics for parameters usually leads to a useful estimation procedure. The discussion up to the present point has been entirely in terms of population distributions. Can we substitute the sample frequencies $f(x)$

for $\phi(x)$ in (1) and (14) or (17), obtaining a useful approximation to $g(\zeta)$? It is found that such a procedure usually produces wholly unusable results even for very large samples. The purpose of the present section is to indicate the nature of the difficulty.

First let us ask, for any given observed-score distribution, $\phi(x)$, is there always a solution--a frequency distribution, $g(\zeta) \geq 0$ --for equation (1)? For simplicity, assume that $g(\zeta)$ is a discrete distribution (this assumption is avoided throughout except at this point) so that (1) can be written

$$\phi(x) = \sum_i g(\zeta_i)\, h(x|\zeta_i) \qquad (x = 0,1,\ldots,n) \,. \tag{18}$$

Let $\Delta^2_\phi(x)$ denote the second difference

$$\Delta^2_\phi(x) \equiv \phi(x) - 2\phi(x-1) + \phi(x-2)$$

and let

$$\Delta^2_i(x) \equiv h(x|\zeta_i) - 2h(x-1 \mid \zeta_i) + h(x-2 \mid \zeta_i) \,.$$

Then from (18)

$$\Delta^2_\phi(x) = \sum_i g(\zeta_i)\, \Delta^2_i(x) \,.$$

For a given value of x and a known set of functions $h(x|\zeta_i)$, $i = 1,2,\ldots$, what is the maximum possible value of $\Delta^2_\phi(x)$?

Denote by $I(x)$ a value of i for which $\Delta^2_i(x)$ takes on its largest value for the given x. Because $g(\zeta)$ is a frequency distribution, $g(\zeta_i) \geq 0$ for all i and $\sum_i g(\zeta_i) = 1$. Consequently, the largest possible $\Delta^2_\phi(x)$ for the given x occurs when $g(\zeta_{I(x)}) = 1$ and all other

$g(\zeta_i) = 0$. In this case, $\Delta^2_\phi(x) = \Delta^2_I(x)$ where $I \equiv I(x)$. Thus for given x , the second difference of $\phi(x)$ can under no circumstances exceed the second difference of $h(x|\zeta_I)$ where $I \equiv I(x)$.

The argument can be extended to apply to continuous $g(\zeta)$. Two conclusions follow. The first is that the actual distribution of test scores is incapable of mirroring any sharp fluctuations that may be present in the distribution of true scores. As was noted in the last section, a very irregular $g(\zeta)$ and a smooth $g(\zeta)$ may give rise to exactly the same, smooth $\phi(x)$.

The second conclusion, for given bounded functions $h(x|\zeta)$, is that <u>if $\phi(x)$ is sufficiently irregular, there can be no distribution $g(\zeta) \geq 0$ satisfying (1)</u>.

In practice, it seems that most sample frequency distributions, $f(x)$, are incompatible with (1). <u>When $f(x)$ is substituted for $\phi(x)$ in (1), any "solution" found for the resulting integral equation usually is irregular and contains negative frequencies</u>.

The following example is not atypical. A 15-item test administered to $N = 3{,}135$ examinees gave the reasonably smooth observed-score distribution shown below. The $\hat{g}(\zeta)$ (estimated $g(\zeta)$) obtained from

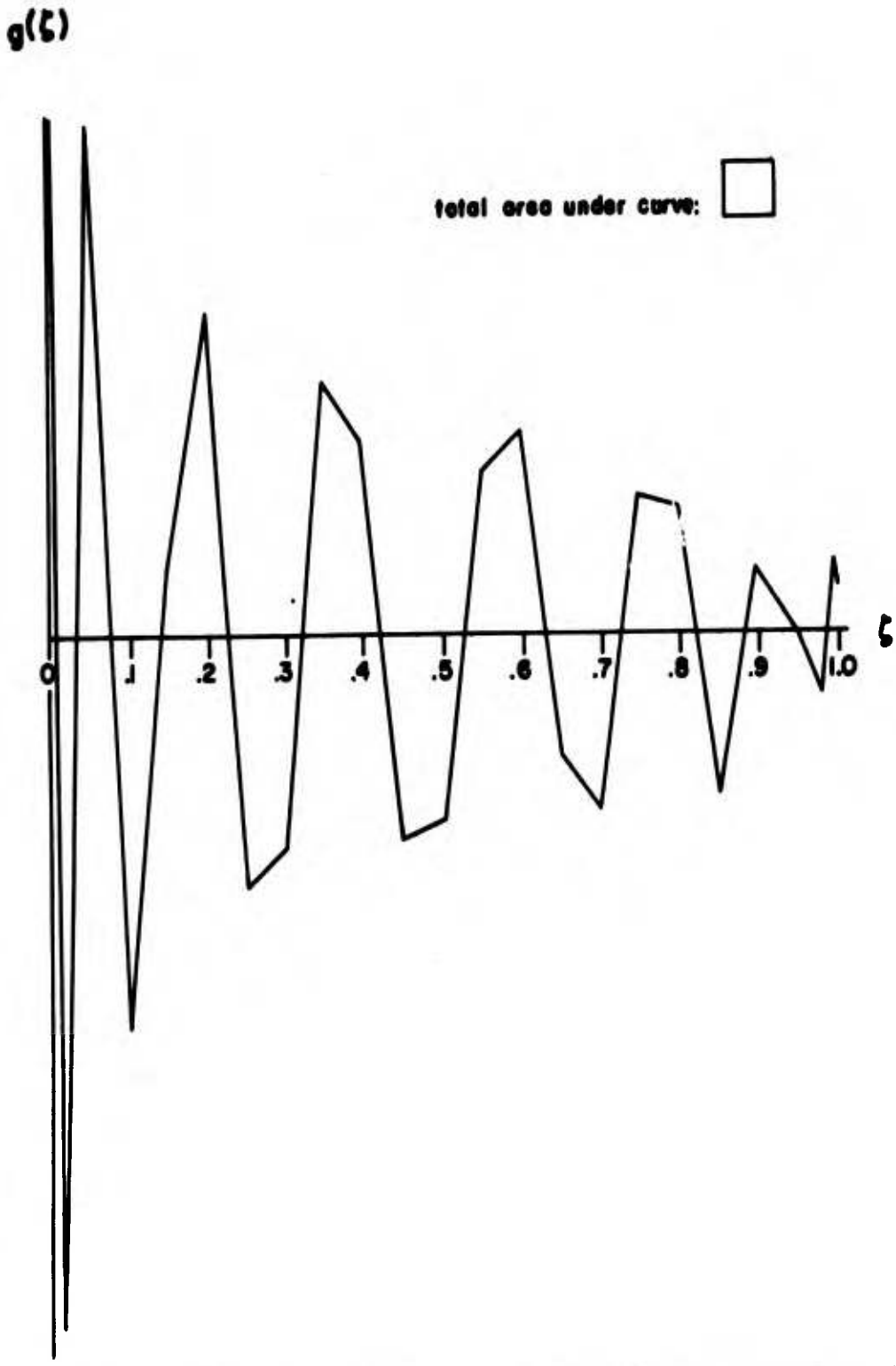


Fig. 1. $g(\zeta)$ giving exact fit to an observed-score distribution in sample of 3,135 examinees.

(14) and (17), with $h(x|\zeta)$ a compound binomial and with $\gamma(\zeta) \equiv 1$, is roughly indicated in Figure 1. Here 25 plotted points have been connected by straight lines in order to avoid the elaborate computations needed to plot the curve accurately. The fluctuations are large; plotted on the same scale, the frequency distribution of observed scores would merge into the horizontal axis and be virtually invisible.

x	Nf(x)
15	163
14	324
13	349
12	363
11	352
10	299
9	276
8	236
7	201
6	189
5	126
4	110
3	88
2	40
1	16
0	3

If a usable estimate, $\hat{g}(\zeta)$, of the true-score distribution can not be found by solving the equation

$$f(x) = \int_a^b \hat{g}(\zeta)\, h(x|\zeta)\, d\zeta \quad ,$$

how then shall $g(\zeta)$ be approximated? This is a problem in statistical inference that has not yet found a widely accepted general solution.

Most statistical estimation is carried out by finding parameters that provide as good a fit to the data as possible within the restrictions imposed by the assumptions made (by the model used). A plausible suggestion is to try to find $\hat{g}(\zeta)$ such that

$$\hat{\phi}(x) \equiv \int_a^b \hat{g}(\zeta)\, h(x|\zeta)\, d\zeta \tag{19}$$

is as close to $f(x)$ as possible in some sense, subject to the restriction that $\hat{g}(\zeta) \geq 0$ for $a \leq \zeta \leq b$. This procedure will get rid of the negative values of $g(\zeta)$ illustrated in Figure 1, but it will not get rid of the numerous peaks. A more complicated procedure involving restrictions on $g(\zeta)$ and also at least on its first two derivatives should be investigated.

6. A Practical Estimation Procedure

A familiar way of dealing with sampling fluctuations in $f(x)$ is to group adjacent values of x and replace $f(x)$ by the corresponding grouped frequency distribution

$$f_u \equiv \sum_{x:u} f(x) \qquad (u = 1,2,\ldots,U)\,, \tag{20}$$

where $\sum_{x:u}$ denotes summation over integers x in the u-th class interval. If $f(x)$ in (1) is replaced by a grouped frequency distribution, and then the method of section 4 with $R = 0$ applied to the grouped distribution, it is readily found that the "smoothest" $g(\zeta)$ is the same as (14) with adjacent λ_x equal for all x in the same group. This smoothest true-score distribution may be written

$$g(\zeta) = \gamma(\zeta) \sum_{u=1}^{U} \lambda_u \sum_{x:u} h(x|\zeta)\,. \tag{21}$$

Here $\gamma(\zeta)$ is a frequency distribution chosen by the statistician (frequently, $\gamma(\zeta) \equiv 1$ --see section 4); the λ_u are U unknown parameters; U is the number of class intervals (groups) in the grouped distribution of x ; and $h(x|\zeta)$ is a known function (in the present case, a compound binomial distribution).

$$g(\zeta) = \sum_{u=1}^{U} \lambda_u H_u(\zeta) \tag{22}$$

where

$$H_u(\zeta) \equiv \gamma(\zeta) \sum_{x:u} h(x|\zeta) \quad .$$

The important function of the grouping is to reduce the number of independent parameters (λ) to be fitted from the data, thus preventing $\hat{g}(\zeta)$ from mirroring too closely irregularities in $f(x)$ due to sampling fluctuations. This is ordinarily necessary in order to prevent $\hat{g}(\zeta)$ from being multimodal.

If (1) is correct, if the number of examinees is large enough, if the grouping is coarse enough, one would expect to find no "negative frequencies" in $\hat{g}(\zeta)$. This may require excessively coarse grouping, however. Experience has shown that a good way to avoid negative $\hat{g}(\zeta)$ is to impose the requirement that

$$\lambda_u \geq 0 \qquad (u = 1,2,\ldots,U) \; . \tag{23}$$

This requirement is much easier to impose than the requirement that

$g(\zeta) \geq 0$ for $a \leq \zeta \leq b$; it is also more effective for reducing multimodality. Although (23) is often more restrictive than necessary, this has not been found to be too serious a problem in actual applications.

If the λ_u were estimated from the grouped distribution f_u using the grouped analogs of equations (15)-(17), then (assuming eq. 23 to be satisfied, or ignored) the estimated grouped distribution defined by

$$\hat{\phi}_u \equiv \sum_{x:u} \hat{\phi}(x) = \int_a^b \hat{g}(\zeta)\, H_u(\zeta)\, d\zeta \qquad (u = 1,2,\ldots,U) \tag{24}$$

would fit the actual grouped distribution exactly; that is, $\hat{\phi}_u$ would equal f_u for all u. The estimation procedure recommended here for the λ_u is an improvement on this: We shall estimate the λ_u by maximum likelihood from the ungrouped distribution $f(x)$ (see Kendall and Stuart, 1958, sections 30.15, 30.19). The restriction (23) is imposed by mathematical programming methods.

The maximum likelihood equations are given in the appendix. The "scoring method" used in their solution is briefly discussed there.

The procedures described in this section, with or without requirement (23), will be referred to as Method 20.

7. Rationales of Analysis

A conventional analysis would estimate $g(\zeta)$ and then evaluate

the adequacy of the model and of the estimation method by a chi square test of significance, comparing the fitted distribution $\hat{\phi}(x)$ of (19) with the actual distribution $f(x)$. It is sometimes objected that since no model is perfect, it is illogical to test statistically the null hypothesis that the model is true. It is urged that statistical significance tests should not be made, that the model should be evaluated according to its practical adequacy (usually as judged subjectively by the reader) rather than by a test of statistical significance.

Here we will consider analyses designed to answer three distinct questions.

1. Does there exist a "smooth" $g(\zeta)$ that under the model would have produced a $\phi(x)$ sufficiently close to the observed $f(x)$? The criterion for "sufficiently close" is not defined, but is left to the reader's practical judgment. Superimposed graphs of $\hat{\phi}(x)$ and $f(x)$ can be shown to aid in this judgment. If no $\hat{g}(\zeta)$ can be found for which $\hat{\phi}(x)$ is sufficiently close to $f(x)$, then one may have to discard the model altogether.

2. Regardless of the answer to the first question, do the present data contain information helpful for modifying and improving the model? If the chi square between $\hat{\phi}(x)$ and $f(x)$ for a particular set of data is near or below the 50-percent level, these data cannot be of much help in improving on the model. This can occur either because the sample is too small or because the model needs little improvement. In order to

$g(\zeta) \geq 0$ for $a \leq \zeta \leq b$; it is also more effective for reducing multimodality. Although (23) is often more restrictive than necessary, this has not been found to be too serious a problem in actual applications.

If the λ_u were estimated from the grouped distribution f_u using the grouped analogs of equations (15)-(17), then (assuming eq. 23 to be satisfied, or ignored) the estimated grouped distribution defined by

$$\hat{\phi}_u \equiv \sum_{x:u} \hat{\phi}(x) = \int_a^b \hat{g}(\zeta)\, H_u(\zeta)\, d\zeta \qquad (u = 1,2,\ldots,U) \tag{24}$$

would fit the actual grouped distribution exactly; that is, $\hat{\phi}_u$ would equal f_u for all u. The estimation procedure recommended here for the λ_u is an improvement on this: We shall estimate the λ_u by maximum likelihood from the ungrouped distribution $f(x)$ (see Kendall and Stuart, 1958, sections 30.15, 30.19). The restriction (23) is imposed by mathematical programming methods.

The maximum likelihood equations are given in the appendix. The "scoring method" used in their solution is briefly discussed there.

The procedures described in this section, with or without requirement (23), will be referred to as Method 20.

7. Rationales of Analysis

A conventional analysis would estimate $g(\zeta)$ and then evaluate

the adequacy of the model and of the estimation method by a chi square test of significance, comparing the fitted distribution $\hat{\phi}(x)$ of (19) with the actual distribution $f(x)$. It is sometimes objected that since no model is perfect, it is illogical to test statistically the null hypothesis that the model is true. It is urged that statistical significance tests should not be made, that the model should be evaluated according to its practical adequacy (usually as judged subjectively by the reader) rather than by a test of statistical significance.

Here we will consider analyses designed to answer three distinct questions.

1. Does there exist a "smooth" $g(\zeta)$ that under the model would have produced a $\phi(x)$ sufficiently close to the observed $f(x)$? The criterion for "sufficiently close" is not defined, but is left to the reader's practical judgment. Superimposed graphs of $\hat{\phi}(x)$ and $f(x)$ can be shown to aid in this judgment. If no $\hat{g}(\zeta)$ can be found for which $\hat{\phi}(x)$ is sufficiently close to $f(x)$, then one may have to discard the model altogether.

2. Regardless of the answer to the first question, do the present data contain information helpful for modifying and improving the model? If the chi square between $\hat{\phi}(x)$ and $f(x)$ for a particular set of data is near or below the 50-percent level, these data cannot be of much help in improving on the model. This can occur either because the sample is too small or because the model needs little improvement. In order to

answer this question, one can compute a chi square and compare its value with its sampling distribution under the model.

3. Are all smooth $g(\zeta)$ that are reasonably consistent with the data much alike? In the population, smooth $g(\zeta)$ consistent with the distribution of observed scores cannot be dissimilar (section 2). This will also be true in sufficiently large samples, but we do not know how large such samples must be. This is a key question, since the answer determines our willingness to accept a smooth $\hat{g}(\zeta)$ consistent with the observed sample as a good approximation to the unknown (presumed smooth) $g(\zeta)$ in the population. All that is done in this direction here is to obtain a variety of $\hat{g}(\zeta)$ from the same data and plot them for visual comparison.

8. Tryout with Hypothetical Data, N = 1000

An estimated true-score distribution $\hat{g}(\zeta)$ may differ from the population value $g(\zeta)$ because of at least four distinct sources of inaccuracy:

1. The mathematical model used here surely falls short of perfection.
2. Since $g(\zeta)$ is not uniquely determined even in the population of examinees and since "smoothness" cannot be uniquely defined, there will be many "smooth" $g(\zeta)$ that satisfy the mathematical model in the population.

3. Sampling fluctuations in the data distort the estimates of $g(\zeta)$.

4. Estimation methods used may fall short of 100 percent efficiency.

The seriousness of the last three sources of error was investigated by generating and analyzing samples of hypothetical data with $N = 1000$. Monte Carlo procedures were used to draw a random sample of $N = 1000$ from the beta distribution

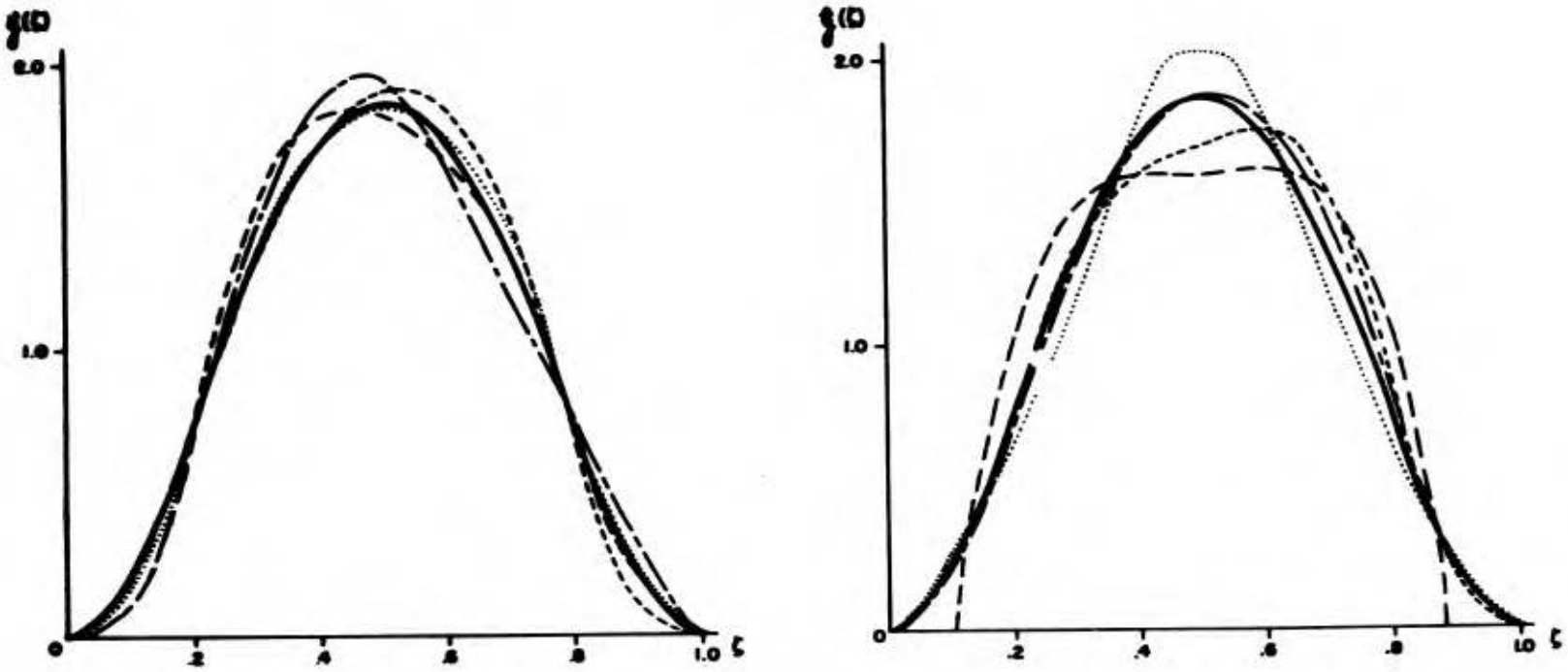
$$g(\zeta) = 30\zeta^2(1 - \zeta)^2 \quad , \qquad 0 \leq \zeta \leq 1 \quad .$$

For each of the 1000 resulting values of ζ , a "raw score" x was drawn independently and at random from the conditional distribution

$$h(x|\zeta) = \binom{n}{x}\zeta^x(1 - \zeta)^{n-x}$$

with $n = 24$. This process was repeated to produce eight independent samples, each representing the observed-score frequency distribution of 1000 hypothetical examinees on a 24-item test.

An estimate $\hat{g}(\zeta)$ was obtained for each of the eight samples using the methods of Lord and Lees (1967b). In every case good agreement was obtained between the fitted observed score distribution $\hat{\phi}(x)$ (equation 19) and the actual $f(x)$. The computed chi squares between $\hat{\phi}(x)$ and $f(x)$ ranged from the 83rd percentile of the chi square sampling distribution (the 17-percent "significance level") down to the 4th percentile.



Figs. 2,3. True-score distribution in a hypothetical population (heavy solid line) and eight estimates of this distribution obtained by Monte Carlo procedures from eight random samples of 1000 cases each.

The estimated true-score distributions obtained from the eight samples are shown in Figures 2 and 3, together with the distribution $g(\zeta) = 30\zeta^2(1 - \zeta)^2$ used to generate the data. Clearly, there exist substantially different smooth $g(\zeta)$ that are consistent with data samples of $N = 1000$ drawn from a single population. (This statement holds true regardless of the fact that the $\hat{g}(\zeta)$ in Figs. 2 and 3 were not obtained by Method 20.) The main conclusion drawn from these and other similar results is that it is desirable to have samples larger than $N = 1000$ if a close approximation to the population $g(\zeta)$ is desired.

9. Description of Tryout Data, N = 20,000

Early versions of Method 20 (described in section 6) were tried out preliminarily on data that had not been well fitted by previously used methods. Method 20 was at least as successful as the others in all cases, and much more successful in some cases.

In view of the results such as those discussed in the preceding section it was decided to use larger groups for the tryout of Method 20 than had been used previously. Four different grade-level groups (grades 4, 6, 8, 10) with about 40,000 pupils each were obtained for study. For each grade, an observed-score distribution was available for each of the

BLANK PAGE

following tests, composed of n four-choice items:

n	Test
50	Mathematics Ability,
50	Mathematics Achievement,
50	Verbal Ability,
30	Reading Achievement,
20	English Achievement.

Grade 10 data for the last two tests were excluded from further study because a much larger number of students scored $x = 0$ than $x = 1$. If there were no guessing, this could occur under the model; but since the tests are composed of four-choice items, such a mode at zero hardly seems plausible. It seems likely that many of the grade 10 examinees who scored 0 really did not attempt the test at all.

Before starting the study, each total group of approximately 40,000 students was split at random into two groups of approximately 20,000 each. This was done separately and independently for each test at each grade level. Trial and error procedures were used on the first-drawn samples of 20,000, designated as the A samples. The B samples were reserved for cross-validation purposes.

10. Results for First-Drawn Samples of 20,000

True-score distributions $\hat{g}(\zeta)$ were estimated by Method 20 for each of the eighteen observed-score distributions studied. Results were eval-

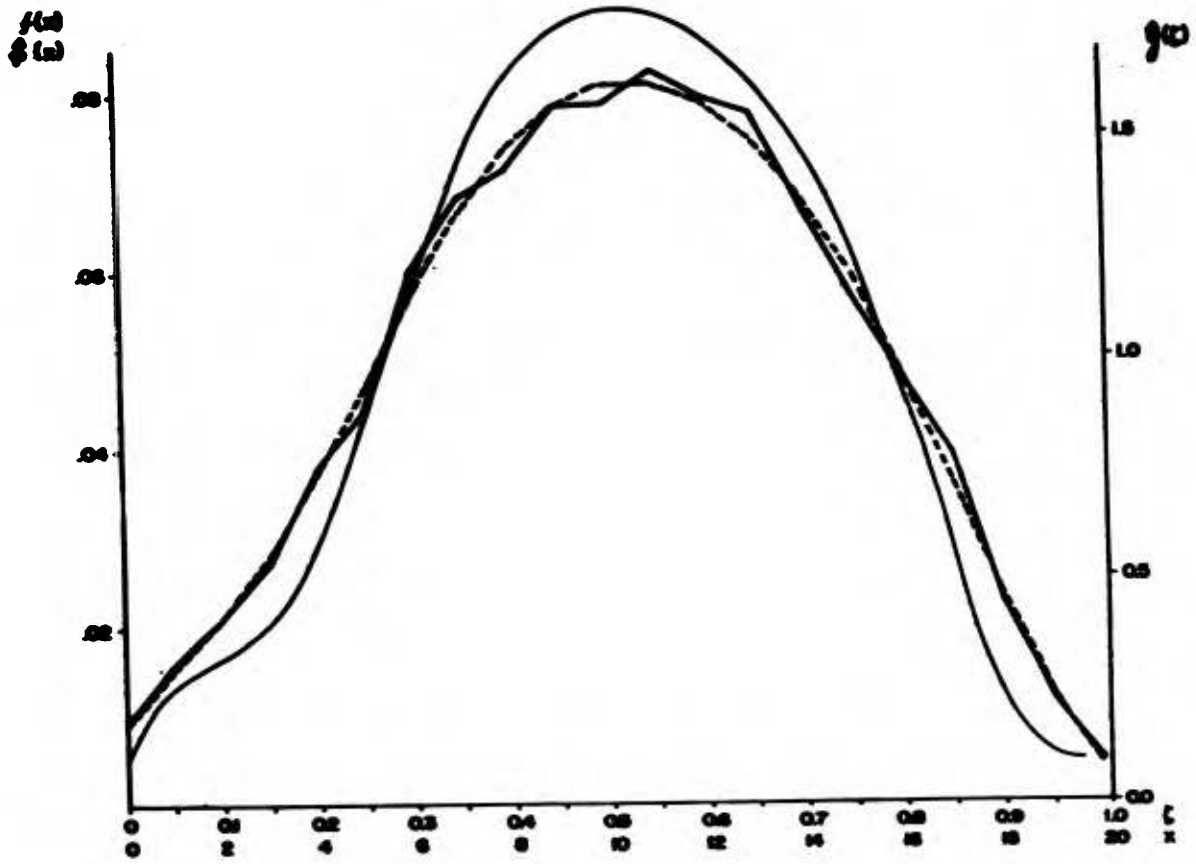


Fig. 4. Observed-score distribution (irregular polygon), fitted observed-score distribution (dashed polygon), and estimated true-score distribution for the worst-fitted set of data in Sample A.

uated initially by computing from $\hat{g}(\zeta)$ the theoretic fitted distribution $\hat{\phi}(x)$ (equation 19) and comparing this with the actual observed-score distribution $f(x)$. If we do not wish to make a chi square test of significance, for reasons outlined in the preceding section, we can make the desired comparison graphically. Instead of presenting all eighteen graphs, Figure 4 presents just the graph for the set of data having the most improbable chi square.

The estimated true-score distribution is shown in the figure for general interest. However, we are mainly concerned with the fit between $\hat{\phi}(x)$ and $f(x)$. Could the reader draw a plausible, smooth $\hat{\phi}(x)$ that would provide a much better fit than the one shown? The fit could be visibly improved near the mode, but this would reduce the chi square only about five percent. More than one-fourth of the total chi square comes from the discrepancy at $x = 17$. It appears that there do exist $g(\zeta)$ that can (under the model) produce observed-score distributions much like those found in the 20,000-case samples.

The $\hat{g}(\zeta)$ used are all represented by equation (22). The number of mathematically independent parameters λ_u is for some data as low as 5 and for some data as high as 12. A more efficient procedure could surely reduce the number of parameters needed for most sets of data. An unnecessarily large number of parameters can often be tolerated when $n = 50$, so that the number of degrees of freedom before fitting is large. It cannot be well tolerated when $n = 20$ and there are not so many degrees of

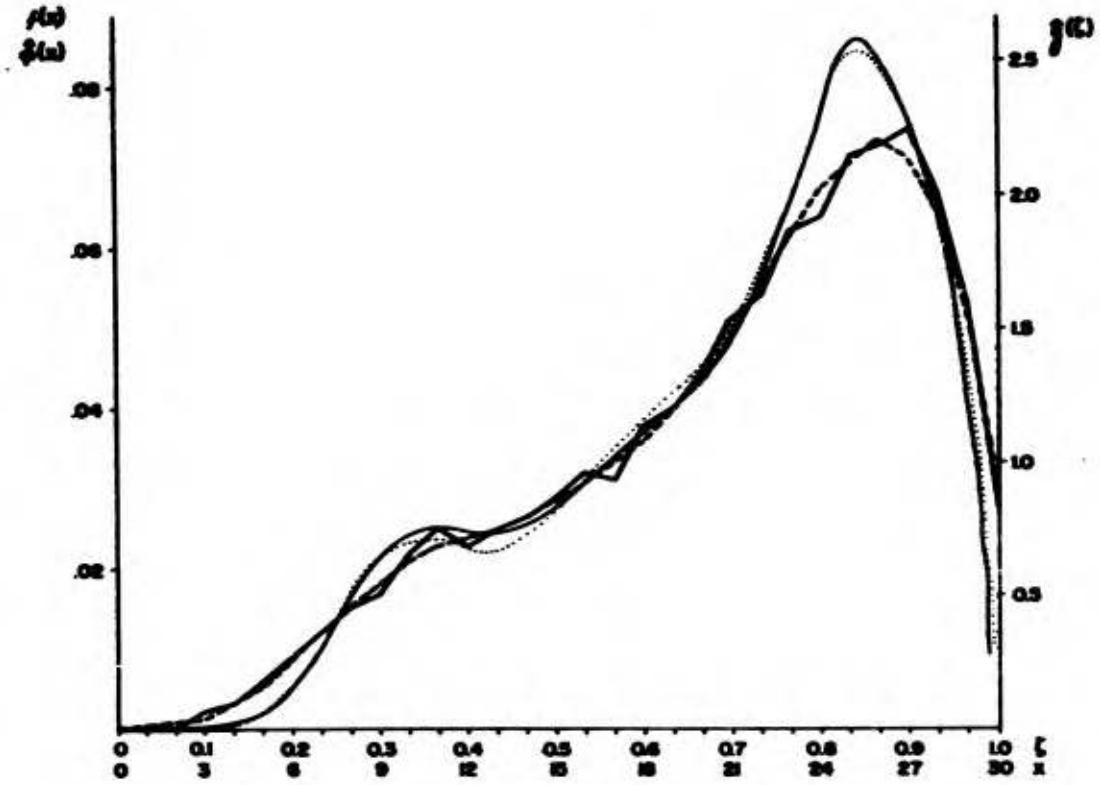


Fig. 5. Observed-score distribution (irregular polygon), fitted observed-score distribution (dashed polygon), and estimated true-score distribution (solid curve) for the next-to-worst-fitted set of data in Sample B; also the same true-score distribution as estimated from Sample A (dotted curve).

freedom to start out with. This fact presumably explains why most of the fitting difficulties occurred with the English $(n = 20)$ and with the Reading $(n = 30)$ tests, whereas very little difficulty was experienced with the Verbal and with the Math tests $(n = 50)$.

The eighteen chi squares found for the eighteen observed-score distributions were (formally) all smaller than the 98th percentile of the sampling distribution of chi squares under the null hypothesis; that is, all were "statistically nonsignificant" at the two-percent level. However, these $\hat{g}(\zeta)$ were obtained by a trial and error procedure that facilitated some capitalization on chance. Thus the obtained chi squares cannot be completely evaluated just by using the theoretical chi square distribution.

11. Results for Cross-Validation Samples of 20,000

In cross-validation, the $\gamma(\zeta)$ and the grouping of the x variable chosen by trial and error in the A sample is used for the corresponding B sample. The values of λ_u are now determined from the B sample. All trial and error takes place on the A sample, none on the B sample.

When this was done, one bad result (grade 6, Reading) was obtained out of the eighteen attempts. Figure 5 shows the next-to-worst result (grade 4, Reading). The fit shown seems good except near the mode. Both grade 4 and grade 6 Reading test scores have rather highly peaked distributions; in both cases the mode of $\hat{\phi}(x)$ is somewhat misplaced for best

fit to the mode of $f(x)$ in the B sample. This difficulty arises because the mode of $f(x)$ in the B sample is not in the same place as in the A sample, and the grouping of the observed scores taken from sample A is incapable of producing a sharp peak where it is needed for sample B.

Good results for both sets of data are obtained simply by using finer grouping of x near the mode. This brings the chi square in the B sample for the grade 6 Reading test down from the 99.9th percentile (.001 "significance level") to the 50th percentile, and the chi square for the grade 4 Reading test down from the 96th percentile to the 88th. Such a use of fine grouping is advantageous in most of the B samples, but it gives rise to a new difficulty; for some data, fine grouping tends to produce undesirably bimodal $\hat{g}(\zeta)$. For this reason, such fine grouping has not been used for the results reported here.

The reader will have noticed that the $\hat{g}(\zeta)$ shown in Figure 5 are bimodal. Since both samples A and B display similar features, both in $\hat{g}(\zeta)$ and in $f(x)$, no attempt was made to prevent this bimodality. A similar situation exists for two other distributions. With these exceptions, no other bimodalities appear in the $\hat{g}(\zeta)$ reported here.

12. Uncertainty in the Estimated True-Score Distribution for Samples of 20,000

When the grouping of the x variable is given, the true-score distribution has a known mathematical form with U unknown parameters

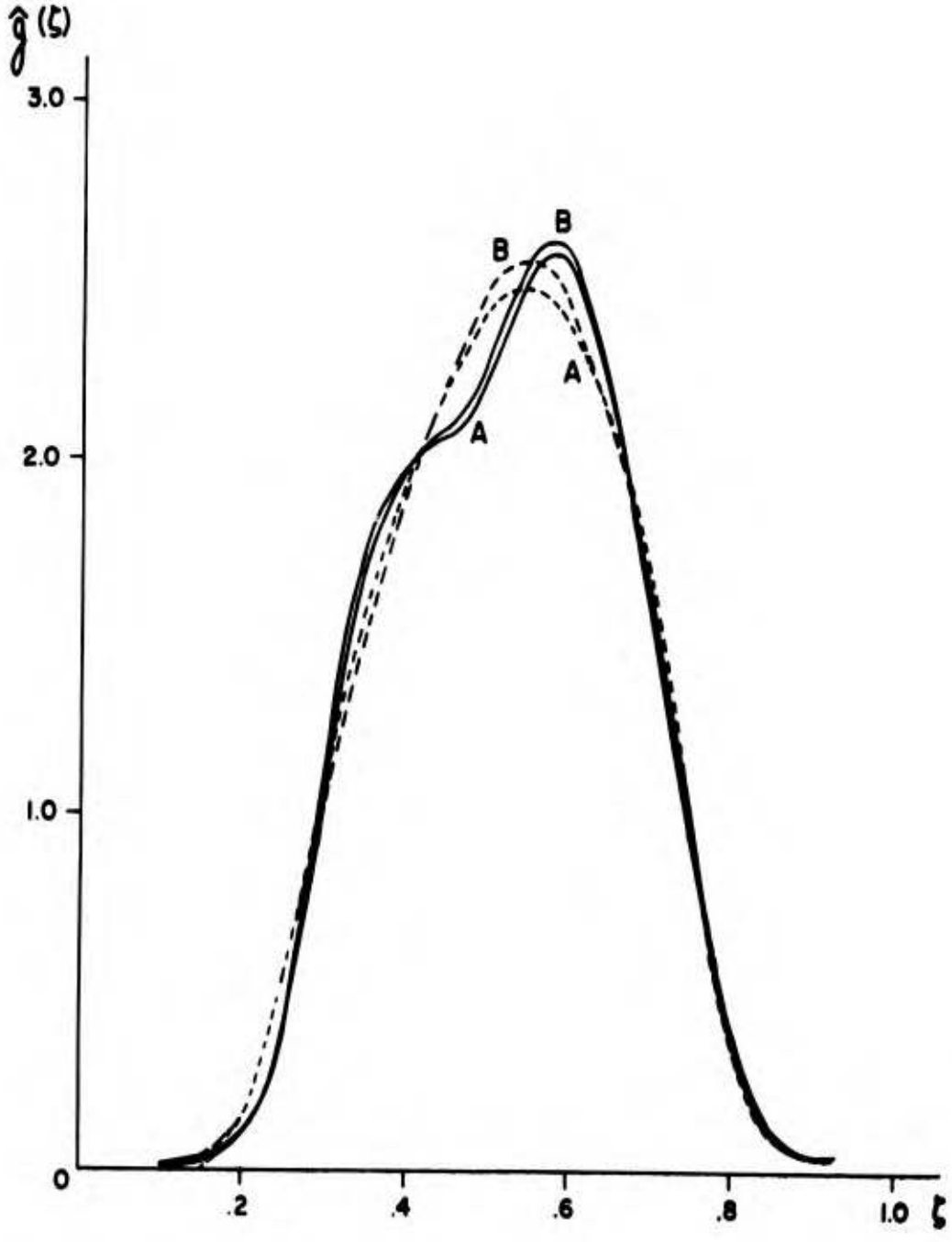


Fig. 6. Estimated true-score distribution obtained for Samples A and B from two groupings of the same data.

(eq. 22), of which U - 1 are mathematically independent of each other (see eq. 31). The various estimated true-score distributions shown in Figures 2 and 3 all have the same mathematical form--the same grouping was used for each distribution. They differ only because of the values assigned to the $U - 1 = 4$ independent parameters. Thus the differences shown there represent sampling fluctuations in the data. They do not indicate what differences might result from different choices of mathematical form for $\hat{g}(\zeta)$, that is, from different groupings of the x variable.

Many different groupings were tried out on a single set of data in the process of trying to fit the sample-A distributions. Commonly, the process was terminated as soon as a good fit was obtained. Occasionally, good fits were obtained for substantially different groupings. Comparisons of the $\hat{g}(\zeta)$ obtained indicate the extent to which substantially different $\hat{g}(\zeta)$ can fit the same set of data.

Usually it was found that any unimodal $\hat{g}(\zeta)$ giving a good chi square between $\hat{\phi}(x)$ and $f(x)$ was much like any other. An exception is shown in Figure 6. Even with 20,000 cases, it is impossible in sample A to choose between the bell-shaped true-score distribution and the "bitangential" true-score distribution. The chi squares are at the 16th and the 63rd percentiles respectively. In sample B, the corresponding chi squares are at the 69th and 98th percentiles.

In both samples, the observed-score distribution shows the same irregularity that tends to produce the bump in the true-score distribution.

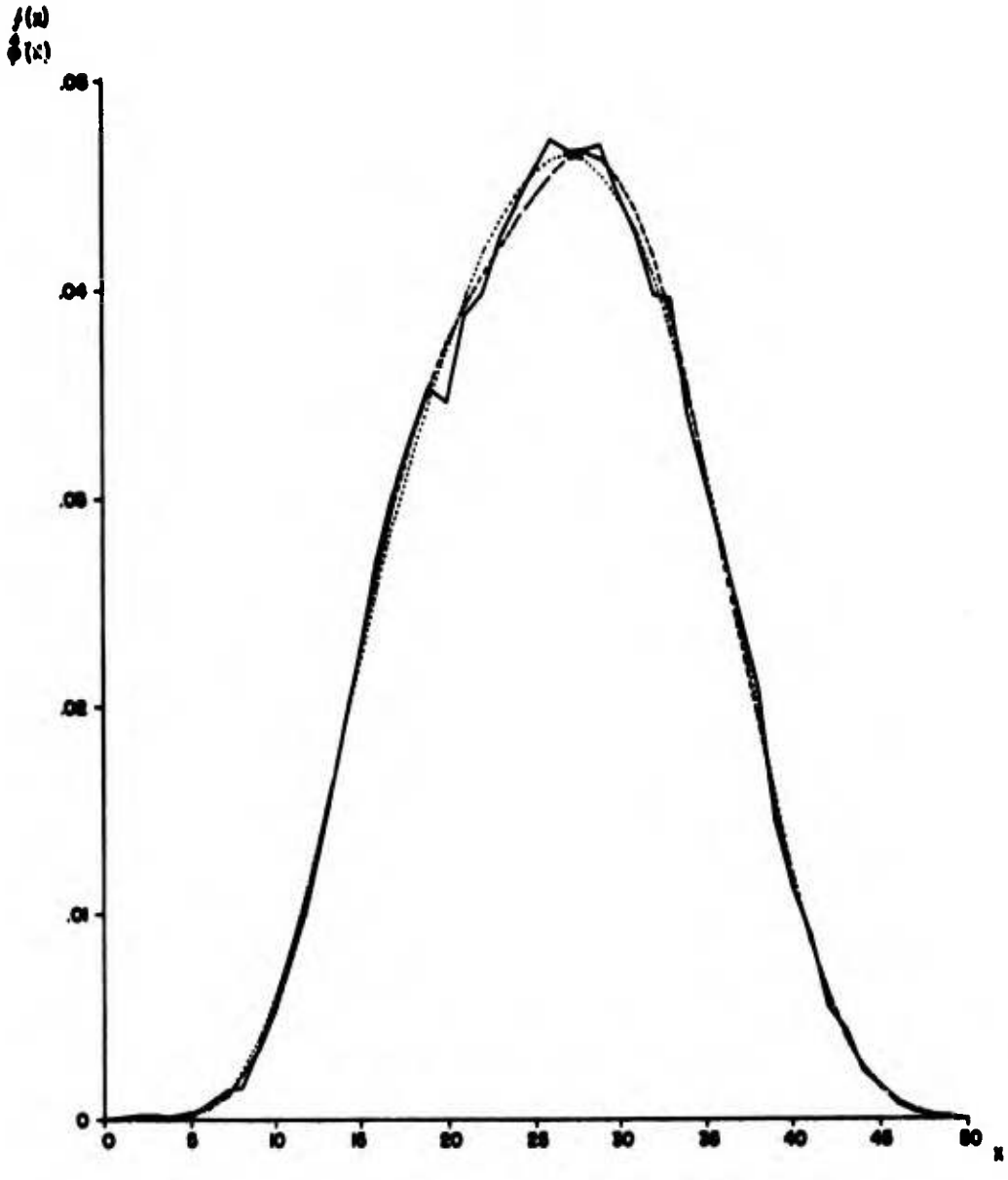


Fig. 7. Observed-score distribution in Sample A and two fitted observed-score distributions computed from the two estimated true-score distributions shown in Figure 6.

The observed-score distribution for sample A appears in Figure 7, along with the two $\hat{\phi}(x)$ generated by the two $\hat{g}(\zeta)$ of Figure 6. It is noteworthy that the $\hat{\phi}(x)$ do not differ as much as do the two $\hat{g}(\zeta)$. This is in line with the ideas developed in section 5.

The conclusion seems to be that even with 20,000 cases we cannot draw firm conclusions about the detailed shape of the true-score distribution. However, in practical applications where the true-score distribution is used only as an intermediate step in computing some characteristic of an observed-score distribution, it may make little difference whether the bell-shaped or the bitangential distribution is used. Just as the bell-shaped and the bitangential distributions give rise to similar $\hat{\phi}(x)$, so also will they give rise to similar estimated bivariate observed-score distributions. It is these last that provide the basis for most practical applications (Lord, 1965).

13. Results for Samples of 200,000

In most work with mathematical models, the larger the sample size, the more likely is the chi square to be significant. In the present problem, there is some indication that Method 20 does not work well if the sample is too small. In order to investigate the effect of sample size, the method was applied to five observed-score distributions with sample sizes ranging from 137,052 to 286,238. No cross-validation samples were used for these sets of data.

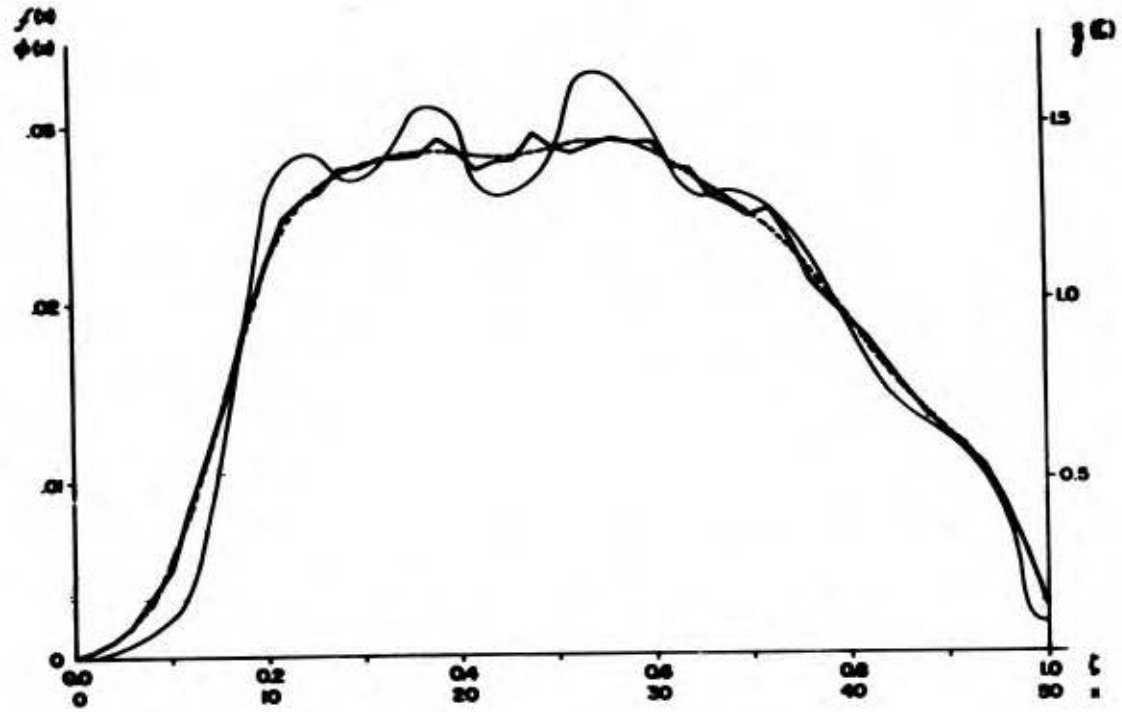


Fig. 8. Observed-score distribution (irregular polygon), fitted observed-score distribution (dashed polygon), and an estimated true-score distribution for the worst-fitted of the five sets of data for which N = 200,000 approximately.

To the writer's surprise, a satisfactory $\hat{g}(\zeta)$ was found for each distribution. The data with the worst chi square are shown in Figure 8. The observed-score distribution has an unusual, flat-topped appearance, which leads to a multimodal $\hat{g}(\zeta)$. The $\hat{g}(\zeta)$ shown has 14 independent parameters. It is a safe assumption that a good fit could be obtained with fewer parameters, and that not all of the modes are needed. However, the necessary effort has not been made to determine how far the $\hat{g}(\zeta)$ can be smoothed without destroying the fit.

It is planned eventually to use a method for finding out how many parameters can be determined without excessive sampling error from a given set of data. This has not yet been worked out in detail.

14. Summary

We have considered the empirical Bayesian model represented by equation (1). In mental test theory, the problem is to estimate the true-score distribution in a population of examinees from the observed-score distribution of a random sample of examinees. The estimated true-score distribution, of interest for its own sake, can be used to draw important practical conclusions about various oberved-score results.

We have assumed (subject to empirical verification) that the conditional distribution of the observed scores for given true score is a certain (approximation to a) compound binomial distribution. If this assumption is correct, then equation (1) can be solved to express the

true-score distribution as a function of the observed-score distribution in the population of examinees. The solution is not unique, but any true-score distribution thus obtained will have the same moments through order n. If we assume that the true-score distribution is "smooth" and without peculiar bulges, then any acceptable solution to (1) cannot differ much from any other.

Formulas for "smooth" solutions to (1) are given for an (infinite) population of examinees. These formulas usually yield absurd results when applied to any sample observed-score distribution. The reason is that the sampling fluctuations in the observed-score distributions reappear in the estimated true-score distribution in greatly magnified form. The problem in dealing with samples is to use some smoothing process while making a minimum of assumptions about the nature of the unknown true-score distribution.

The procedure suggested here starts with the formula (equation 14) for estimating the population true-score distribution under the assumption that it is smooth in a certain specified sense. Smoothing is achieved by replacing the ungrouped observed-score frequencies by grouped observed-score frequencies in this formula. The parameters of the formula are then estimated from the ungrouped data by maximum likelihood, subject to certain inequalities on the estimates that prevent the occurrence of "negative frequencies" in the estimated true-score distribution.

Certain characteristics of the estimation method may be listed:

1. It makes a minimum of assumptions about the nature of $g(\zeta)$. (Although one might expect that almost any reasonable four-parameter frequency distribution would represent $g(\zeta)$ adequately for present purposes, several years of experience with various assumed forms for $g(\zeta)$ has shown that this is definitely not the case.)
2. The choice among the many different mathematical forms available in this method is made on the basis of the data. Thus the method capitalizes on chance to a significant extent.
3. This can be dealt with by splitting the data into two random samples, using one sample to choose the mathematical form of $\hat{g}(\zeta)$, and then checking the adequacy of the procedure on the other sample. For the data reported here, the increase in chi square in the cross-validation sample has been rather small for most (but not all) sets of data.
4. In samples of $N = 1000$, the method may lead to a bumpy $\hat{g}(\zeta)$. Although such bumps can often be avoided by repeated trial-and-error procedures, it is more satisfactory to start with a larger sized

sample ($N > 10{,}000$, say).

5. Although relatively convenient procedures have been worked out for selecting a mathematical form for $\hat{g}(\zeta)$ on the basis of the observed data, the procedures are inefficient, leading to use of a $\hat{g}(\zeta)$ having more free parameters to be determined from the data than are really needed. This represents a loss in degrees of freedom that is more easily tolerated in work with longer tests ($n \geq 40$) than in work with shorter tests.

6. The procedure seems to be capable of fitting satisfactorily most univariate observed-score distributions.

It would be desirable to check out the effectiveness of this model for estimating the bivariate distribution of observed scores for two tests measuring the same psychological trait. This has not yet been done because of the difficulty of obtaining suitable scatterplots with a sufficiently large N. It has been found in the past (Lord, 1965; Lord & Lees, 1967a, 1967b), however, that such bivariate distributions are usually fitted more readily than are their univariate marginal distributions.

In the writer's opinion, the method appears to be ready for practical use in norming and other problems, at least where $N > 10{,}000$. The ability of the method to predict bivariate observed-score distributions for two tests measuring the same trait should be checked in the course of any such application.

APPENDIX

Equation (21) gives the $g(\zeta)$ that fits the grouped observed-score distribution and is smoothest in the (very limited) sense of (11) and (12). As pointed out in section 4, the "smoothing function" $\gamma(\zeta)$ can usually be chosen to be $\gamma(\zeta) \equiv 1$ without drastic effect on the $g(\zeta)$ obtained.

If the $U - 1$ mathematically independent parameters λ_u were determined from the $U - 1$ mathematically independent grouped frequencies $f_u \equiv \sum_{x:u} f(x)$, the fitted frequencies $\hat{\phi}(u)$ obtained from (24) would fit the grouped frequencies exactly. We modify this procedure in two respects:

1. We estimate the λ_u using the full information provided by the ungrouped frequencies (see Kendall and Stuart, 1958, sect. 30.15, 30.19).
2. We require that each λ_u be nonnegative.

We start with a random sample of observations $x_1, x_2, \ldots, x_N$ drawn from the frequency distribution

$$(25) \qquad \phi(x) = \int_a^b g(\zeta)\, h(x|\zeta)\, d\zeta \qquad (x = 0,1,\ldots,n)$$

$$= \sum_{u=1}^{U} a_{xu} \lambda_u \quad ,$$

where

$$a_{xu} = \sum_{x':u} \int_a^b \gamma(\zeta)\, h(x'|\zeta)\, h(x|\zeta)\, d\zeta \qquad (x = 0,1,\ldots,n;\ u = 1,2,\ldots,U) \quad . \tag{26}$$

Equations (25) and (26) are obtained by substituting (22) into (1).

Suppose $\gamma(\zeta)$ is chosen to have the form of a two-parameter beta distribution:

$$\gamma(\zeta) \propto \zeta^{d}(1 - \zeta)^{\Delta} \quad , \tag{27}$$

where d and Δ are chosen for convenience to be integers. The usual case where $\gamma(\zeta) =$ constant is the special case of (27) where $d = \Delta = 0$. If $h(x|\zeta)$ is binomial, then the integral in (26) is proportional to

$$\int_a^b \zeta^{d+x'+x} (1 - \zeta)^{\Delta+2n-x'-x}\, d\zeta \quad . \tag{28}$$

Since the exponents are integers, this integral can be evaluated precisely for successive values of x' and x by a convenient recursive procedure (Jordan, 1947, sect. 25, eq. 5).

If $h(x|\zeta)$ is the series approximation to a compound binomial distribution used by Lord and Lees (1967a), then each integral in (26) is itself a simple weighted sum of terms like (28). The weights used for the two-term series approximation are given in Lord (1965, eq. 56,57);

for the four-term approximation, in Lord and Lees (1967a, eq. 56).

When the (n+1) -by- U matrix of numerical values a_{xu} has been computed, the only remaining problem is to estimate the parameters λ_1 , λ_2 ,..., λ_U of the distribution $\phi(x)$ from the sample of observations x_1 , x_2 ,..., x_N . The estimated values of λ can then be inserted into (22) to obtain $\hat{g}(\zeta)$, the estimated true-score distribution. When $\gamma(\zeta)$ is chosen to be a beta distribution, then $\hat{g}(\zeta)$ is seen to be a polynomial of degree $n + d + \Delta$.

A1. Maximum Likelihood Estimation

The sample frequencies f_0 , f_1 ,..., f_n , where $\Sigma_x f_x = N$, have a multinomial distribution proportional to

$$L = \prod_{x=0}^{n} [\phi(x)]^{f_x} \quad . \tag{29}$$

Thus the logarithm of the likelihood function is

$$\log L = f_0 \log \phi(0) + f_1 \log \phi(1) + \ldots + f_n \log \phi(n) + \text{constant} \quad . \tag{30}$$

It is effective to use Fisher's scoring method (see Rao, 1965, Chap. 5g, example 3) to find the values $\hat{\lambda}_1$, $\hat{\lambda}_2$,..., $\hat{\lambda}_U$ that maximize (30).

Before proceeding, note that when we sum (25) on x , we find that

$$(31) \qquad \sum_{u=1}^{U} A_u \lambda_u = \sum_{x=0}^{n} \phi(x) = 1$$

where

$$(32) \qquad A_u = \sum_{x=0}^{n} a_{xu} \qquad (u = 1,2,\ldots,U) .$$

Thus there is a restriction on the λ_u : only $U - 1$ of the λ_u are mathematically independent. The remaining parameter is determined from the others by use of (31).

We will need the <u>efficient scores</u>

$$S_u = \frac{\partial \log L}{\partial \lambda_u} ,$$

for any $U - 1$ of the λ_u . From (30) and (25),

$$(33) \qquad \log L = \sum_{x=0}^{n} f_x \log \sum_{u=1}^{U} a_{xu} \lambda_u + \text{constant} .$$

Let us treat the first $U - 1$ of the λ_u as the independent parameters to be determined.

From (25) and (31),

$$(34) \qquad \phi'_u(x) = \frac{\partial}{\partial \lambda_u} \phi(x) = a_{xu} + \frac{\partial \phi(x)}{\partial \lambda_U} \frac{\partial \lambda_U}{\partial \lambda_u} = a_{xu} - \frac{A_u}{A_U} a_{xU} .$$

The value of S_u is then (see Rao) conveniently obtained from

$$S_u = \sum_{x=0}^{n} f_x \frac{\phi_u'(x)}{\phi(x)} \quad . \tag{35}$$

The maximum likelihood estimates $\hat{\lambda}_u$, $u = 1,2,\ldots,U-1$, are the solutions to the equations $S_u = 0$, $u = 1,2,\ldots,U-1$. The iterative procedure for solving these equations requires the information matrix whose elements are

$$I_{uv} = N \sum_{x=0}^{n} \frac{\phi_u'(x)\,\phi_v'(x)}{\phi(x)} \quad . \tag{36}$$

The scoring procedure proceeds by solving for $\Delta_1^{(r)}, \Delta_2^{(r)},\ldots, \Delta_{U-1}^{(r)}$ the linear equations

$$\sum_{v=1}^{U-1} I_{uv}^{(r)} \Delta_v^{(r)} = S_u^{(r)} \qquad (u = 1,2,\ldots,U-1), \tag{37}$$

where $S_u^{(r)}$ denotes the r-th trial value. New trial values are obtained from

$$\hat{\lambda}_u^{(r+1)} = \hat{\lambda}_u^{(r)} + \Delta_u^{(r)} \quad . \tag{38}$$

In the present case, initial trial values $\hat{\lambda}_u^{(1)}$ were obtained by solving the U linear equations

$$(39) \qquad f_u \equiv \sum_{x:u} f(x) = \sum_{v=1}^{U} \hat{\lambda}_v^{(1)} \sum_{x:u} a_{xu} \qquad (u = 1,2,\ldots,U) \quad .$$

With this starting point, the iterations were found in practice to converge in practically every case tried.

The estimated true-score distribution is now obtained by substituting $\hat{\lambda}$ for λ in (22). Thus values of $\hat{g}(\zeta)$ are computed from

$$(40) \qquad \hat{g}(\zeta) = \sum_{u=1}^{U} \hat{\lambda}_u H_u(\zeta) \quad .$$

A2. Restricted Maximum Likelihood Estimation

The following procedure was used to keep $\hat{\lambda}_u^{(r)} \geq 0$. First of all, any negative initial trial values $\hat{\lambda}_u^{(1)}$ obtained from (39) were replaced by a small positive constant, after which all values of $\hat{\lambda}_u^{(1)}$ were decreased proportionately to satisfy equation (31).

The scoring procedure was then applied to these nonnegative $\hat{\lambda}_u^{(1)}$. If at any stage the scoring procedure produced negative trial values $\hat{\lambda}_u^{(r+1)}$, these were discarded and instead all values of $\Delta_u^{(r+1)}$ were decreased proportionately in absolute value, just enough so that one $\hat{\lambda}_u^{(r+1)}$ computed from (38) was exactly zero while all others were nonnegative.

Suppose that $\hat{\lambda}_{u_o}^{(r+1)}$ becomes exactly zero in this way. It is then temporarily assumed that λ_{u_o} is known to be zero. Thus (25) is replaced by

$$\phi(x) = \sum_{u \neq u_o} a_{xu} \lambda_u \tag{41}$$

and the number of unknown parameters to be estimated is reduced by one. The obvious modifications of the likelihood function and of the scoring procedure corresponding to (41) are made and the iterative process is continued with one less unknown parameter.

In the course of repetition of the foregoing, several λ are likely to be fixed at zero. The iterative process finally converges, assigning positive values to the remaining λ .

At this point, one λ , previously fixed at zero, is reintroduced as a parameter to be estimated along with the currently nonzero parameters. The whole process already described is repeated. This reintroduction of parameters is systematically continued until a point $(\lambda_1, \lambda_2, \ldots, \lambda_U)$ is reached such that the likelihood function (29) cannot be increased by any small change in any λ , within the restriction that $\lambda_u \geq 0$. Such a point gives a restricted maximum of the likelihood function.

It has not been shown that such a maximum is a global rather than merely a local maximum. In practice, it has been found that when the iterative process is repeated starting with different initial trial values, the same maximum is found again.

A3. Estimating the Parameters a and b

The range of the true score is $0 \leq a \leq \zeta \leq 1$. Ideally it should be possible to set $a = 0$ and $b = 1$. Any true score with an effective range shorter than [0,1] would simply have $g(\zeta) = 0$ outside its effective range. In practice, it is usually preferable when possible to set $\hat{a} = .02$ and $\hat{b} = .98$ since the four-term approximation to the compound binomial may produce a slightly negative $h(x|\zeta)$ for certain values of x when ζ is too extreme.

If no restrictions are placed on the λ, it is sometimes found that use of $\hat{a} = 0$ or $\hat{a} = .02$ and $\hat{b} = 1$ or $\hat{b} = .98$ leads to negative values in the tails of $\hat{g}(\zeta)$. This difficulty can frequently be avoided by choosing less extreme values for $\hat{a}$ and $\hat{b}$. A trial and error process usually seems quite adequate for this purpose. The values of a and b could be estimated by maximum likelihood, if desired.

When the λ are required to be nonnegative, the values $a = 0$, $b = 1$ or $a = .02$, $b = .98$ usually seem to be satisfactory.

A4. The Sampling Variance of the Estimated True-Score Distribution

When the λ are unrestricted, the variance-covariance matrix of any set of $U - 1$ estimators $\hat{\lambda}_u$ is the inverse of the corresponding

Table 1

Sampling Error of Estimated True-Score Distribution

at Various Levels of True Score

ζ	$\hat{g}(\zeta)$	$\sqrt{\text{Var}[\hat{g}(\zeta)\mid\zeta]}$
.95	.6	.035
.85	1.5	.041
.75	1.7	.032
.65	1.6	.029
.55	1.3	.033
.45	1.2	.028
.35	1.1	.032
.25	0.7	.018
.15	0.2	.022
.05	0.1	.014

matrix $\|\hat{I}_{uv}\|$. This last is automatically computed by (36) as part of the scoring procedure.

By (31)

$$\hat{\lambda}_U = 1 - \frac{1}{A_U} \sum_{u=1}^{U-1} A_u \hat{\lambda}_u \quad . \tag{42}$$

Thus

$$\operatorname{Var} \hat{\lambda}_U = \frac{1}{A_U^2} \sum_{u=1}^{U-1} \sum_{v=1}^{U-1} A_u A_v \operatorname{Cov}(\hat{\lambda}_u, \hat{\lambda}_v) \quad , \tag{43}$$

$$\operatorname{Cov}(\hat{\lambda}_U, \hat{\lambda}_u) = - \frac{1}{A_U} \sum_{v=1}^{U-1} A_v \operatorname{Cov}(\hat{\lambda}_u, \hat{\lambda}_v) \qquad (u = 1,2,\ldots,U-1) \quad . \tag{44}$$

From this, the variance-covariance matrix of all U estimators can be written down.

The sampling variance of $\hat{g}(\zeta)$ for any fixed value of ζ can now be written down from (40), at least for the case where it is permitted that $\lambda_u < 0$:

$$\operatorname{Var}[\hat{g}(\zeta) | \zeta] = \sum_{u=1}^{U} \sum_{v=1}^{U} H_u(\zeta) H_v(\zeta) \operatorname{Cov}(\lambda_u, \lambda_v) \quad . \tag{45}$$

The sampling variance of $\hat{g}(\zeta)$ has been computed for various values of ζ for many of the distributions studied. The results for one distribution are listed in Table 1. The standard errors found are too small to justify a graphical presentation. Such sampling variances should not be confused with discrepancies arising from different choices of functional form for $\hat{g}(\zeta)$ -- that is, from differently chosen groupings of the observed-score variable.

References

Jordan, C. Calculus of finite differences. (2nd ed.) New York: Chelsea, 1947.

Kendall, M. G., & Stuart, A. The advanced theory of statistics. New York: Hafner, 1958-61, 2 volumes.

Kenneth, P., & Taylor, G. E. Solution of variational problems with bounded control variables by means of the generalized Newton-Raphson method. In A. Lavi & T. Vogl (Eds.), Recent advances in optimization techniques. New York: Wiley, 1966.

Leitmann, G. Variational problems with bounded control variables. In G. Leitmann (Ed.), Optimization techniques. New York: Academic Press, 1962. Pp. 171-204.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M., & Lees, Diana. Estimating true-score distributions for mental tests (Methods 12, 14, 15). ETS Research Memorandum 67-1 and ONR Technical Report, Contract Nonr 2752(00). Princeton, N. J.: Educational Test ervice, 1967. (a)

Lord, F. M., & Lees, Diana. Estimating true-score distributions for mental tests (Method 16). ETS Research Bulletin 67-7 and ONR Technical Report, Contract Nonr 2752(00). Princeton, N. J.: Educational Testing Service, 1967. (b)

Maritz, J. S. Smooth empirical Bayes estimation for one-parameter discrete distributions. Biometrika, 1966, 53, 417-429.

Pars, L. A. An introduction to the calculus of variations. London: Heinemann, 1962.

Rao, C. R. Linear statistical inference and its applications. New York: Wiley, 1965.

Riordan, J. An introduction to combinatorial analysis. New York: Wiley, 1958.

Robbins, H. The empirical Bayes approach to statistical decision problems. The Annals of Mathematical Statistics, 1964, 35, 1-20.

Skellam, J. G. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between sets of trials. Journal of the Royal Statistical Society, Series B, 1948, 10, 257-261.

Tricomi, F. G. Integral equations. New York: Interscience, 1957.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Educational Testing Service Princeton, New Jersey 08540	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

ESTIMATING TRUE-SCORE DISTRIBUTIONS IN PSYCHOLOGICAL TESTING (AN EMPIRICAL BAYES ESTIMATION PROBLEM)

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

Frederic M. Lord

6. REPORT DATE November 1967	7a. TOTAL NO. OF PAGES 53	7b. NO. OF REFS 14
8a. CONTRACT OR GRANT NO. Nonr 2752(00) b. PROJECT NO. NR 151-201 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) RB-67-37 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Office of Naval Research Navy Department Washington, D. C. 20360

13. ABSTRACT

The following problem is considered: Given that the frequency distribution of the errors of measurement is known, determine or estimate the distribution of true scores from the distribution of observed scores for a group of examinees. Typically this problem does not have a unique solution. However, if the true-score distribution is "smooth," then any two smooth solutions to the problem will differ little from each other. Methods for finding smooth solutions are developed a) for a population and b) for a sample of examinees. The results of a number of tryouts on actual test data are summarized.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

Security Classification

SND PPSO 13152

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Mental Test Theory						
Psychological Examinations						
Empirical Bayes Estimation						