Practical Applications of Item Characteristic Curve Theory

Author(s): Frederic M. Lord

Source: *Journal of Educational Measurement*, Summer, 1977, Vol. 14, No. 2, Applications of Latent Trait Models (Summer, 1977), pp. 117-138

Published by: National Council on Measurement in Education

Stable URL: https://www.jstor.org/stable/1434011

JOURNAL OF EDUCATIONAL MEASUREMENT
Volume 14, No. 2 Summer 1977

PRACTICAL APPLICATIONS OF ITEM CHARACTERISTIC CURVE THEORY*

FREDERIC M. LORD
Educational Testing Service

Much of classical test theory deals with an entire test; this theory is applicable even if the test is not composed of items. If a test consists of separate items and if test score is a (possibly weighted) sum of item scores, then statistics describing the test scores of a certain group of examinees can be expressed algebraically in terms of statistics describing the individual item scores for the same group of examinees. Insofar as it relates to tests, classical item theory (this is only a part of classical test theory) consists of such algebraic tautologies. Such a theory makes no assumptions about matters that are beyond the control of the psychometrician. This is actuarial science. It cannot predict how individuals will respond to items unless the items have previously been administered to similar individuals.

In practical test development work, we often need to predict the statistical properties of a test composed of items for a target group of examinees that is somewhat different from the groups to which the separate items have previously been administered. We need to be able to describe the items by using item parameters, and the examinees by using examinee parameters, in such a way that we can predict probabilistically the response of any examinee to any item, even if similar examinees have never taken similar items before.

This involves making predictions about things beyond the control of the psychometrician—about how people will behave in the real world. Thus some assumption must be made (and verified) as to how a person's ability or skill determines hir performance on items measuring that ability or skill. Item response theory (item characteristic curve theory) provides these assumptions.

ESTIMATING THE STATISTICAL CHARACTERISTICS OF A TEST FOR ANY SPECIFIED GROUP

How can item response theory be applied? It is sometimes asserted that item response theory allows us to answer any question that we are entitled to ask about the characteristics predicted for a test composed of items with known item parameters. The significance of this vague statement arises from the fact that item response theory provides us with the frequency distribution f of test score for examinees having a specified level θ of ability or skill.

Except where otherwise noted, consideration here will be limited to the number-right score, denoted by x. Suppose the n items in a test all had identical item response curves $P \equiv P(\theta)$ (item characteristic curves, see Hambleton & Cook (1977) for a detailed definition). The distribution of x for a person at ability level θ would then be the familiar binominal distribution $f(x \mid \theta) = \binom{n}{x} P^x Q^{n-x}$, where $Q \equiv 1 - P$. The

*This paper includes portions of the writer's 1976 presidential address to Division 5 of APA, titled "Applications of item response theory to practical testing problems." Also portions of the writer's talk, titled "Test theory in the public interest," at the 1976 ETS Invitational Conference.

expression $(Q + P)^n$ is familiar as the generating function for the binomial distribution because the binomial expansion

$$(Q + P)^n \equiv Q^n + nPQ^{n-1} + \binom{n}{2}P^2Q^{n-2} + \cdots \binom{n}{x}P^xQ^{n-x} + \cdots + P^n$$

gives the terms of $f(x \mid \theta)$ successively for $x = 0, 1, \ldots, n$. [These and other formulas are given here to suggest the nature of the results obtained; no mathematical manipulations are contemplated for this presentation.]

When the item response curves $P_i \equiv P_i(\theta)$ $(i = 1, 2, \ldots, n)$ vary from item to item, as is ordinarily the case, $f(x \mid \theta)$, the frequency distribution of the number-right test score for given θ, is a generalized binomial. This distribution can be generated by using the generating function

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

For example, if $n = 3$, the computer computes the four frequencies $Q_1Q_2Q_3$, $Q_1Q_2P_3 + Q_1P_2Q_3 + P_1Q_2Q_3$, $Q_1P_2P_3 + P_1Q_2P_3 + P_1P_2Q_3$, and $P_1P_2P_3$ by a convenient recursive procedure.

Although $f(x \mid \theta)$, the distribution of number-right scores, cannot be simply written explicitly, the mean $\mu_{x \mid \theta}$ and the variance of $\sigma^2_{x \mid \theta}$ for given θ are simply

$$\mu_{x \mid \theta} = \sum_{i=1}^{n} P_i, \tag{2}$$

and

$$\sigma^2_{x \mid \theta} = \sum_{i=1}^{n} P_iQ_i. \tag{3}$$

Given the functional form of $P_i(\theta)$ (see Hambleton & Cook) and given the parameters of n items, we can explicitly calculate $f(x \mid \theta)$ using Equation (1) for any θ. Given the ability levels $\theta_1, \theta_2, \ldots, \theta_N$ of any group of N examinees, we can predict the frequency distribution of number-right score that will actually be observed when the n-item test is administered to this group of examinees. The estimate is simply

$$\hat{f}(x) = \frac{1}{N}\sum_{a=1}^{N} f(x \mid \theta_a). \tag{4}$$

Such predicted distributions are readily obtained using a fast computer, even for $n = 100$ and $N = 3000$.

The predicted mean score for the group may be computed from (compare Equation 2)

$$\hat{\mu}_x = \frac{1}{N}\sum_{a=1}^{N}\sum_{i=1}^{n} P_{ia}, \tag{5}$$

where $P_{ia} \equiv P_i(\theta_a)$. The predicted score variance $\hat{\sigma}^2_x$ for the group can be obtained from the P_{ia} by formula, or can be computed directly from $\hat{f}(x)$ in Equation (4).

The squared (average) standard error of measurement of classical test theory, $\sigma^2_{x\cdot t}$, for the total group can be estimated conveniently from

$$\hat{\sigma}^2_{x\cdot t} = \frac{1}{N}\sum_{a=1}^{N}\sum_{i=1}^{n} P_{ia}Q_{ia} \tag{6}$$

(compare Equation 3). Here t denotes true score. The classical test reliability coefficient $\rho_{xx'}$ can then be estimated from

$$\hat{\rho}_{xx'} = 1 - \hat{\sigma}^2_{x\cdot t}/\hat{\sigma}^2_x. \tag{7}$$

All this means that if we have a pool of pretested items, all measuring the same trait or ability, we can predict the mean, variance, reliability and raw-score frequency distribution of any test constructed from these items once we know the ability levels in the group to be tested. This effectively illustrates some of the capabilities of item response theory.

SELECTING ITEMS FOR A CONVENTIONAL TEST

For any method of scoring a test, the information function for the score is defined by

$$I(\theta, y) \equiv \frac{(d\mu_{y|\theta}/d\theta)^2}{\sigma^2_{y|\theta}}. \tag{8}$$

This definition does not even require that the test be composed of 'items,' only that y be some score, however complicated, proposed as a measure of θ. This is not a computing formula. When mathematical formulas for $\mu_{y|\theta}$ and $\sigma_{y|\theta}$ are known, as in Equations (2) and (3) for number-right scores, the derivative and the denominator in Equation (8) are replaced by specific functions of $P(\theta)$, after which $I(\theta, y)$ can be computed from $P(\theta)$ for any given θ.

For a conventional test, if $y \equiv \hat{\theta}$, the maximum likelihood estimator of θ, it is found from Equation (8) that

$$I(\theta, \hat{\theta}) = \sum_{i=1}^{n} P_i'^2/P_iQ_i \tag{9}$$

where $P_i' \equiv dP_i(\theta)/d\theta$. Again, the derivative $P'(\theta)$ can be replaced by a specific formula once the form of $P(\theta)$ is specified. For any given θ, $I(\theta, \hat{\theta})$ is then readily computed from the item parameters. Thus Equation (9) can be computed without computing $\hat{\theta}$—a value which can only be computed iteratively.

For any scoring method y, $I(\theta, \hat{\theta})$ is an upper bound for $I(\theta, y)$. In practice, $I(\theta, \hat{\theta})$ is usually a good approximation to $I(\theta, y)$, except perhaps at low ability levels. If there is much difference between the two, a better score than y should be found for scoring the test (see the section on optimal scoring weights).

A striking feature of Equation (9) is that the contribution of each item to this test information function can be determined without knowing what other items are in the test. When building a test, this is invaluable for deciding which items to include. In classical test theory, the situation is very different. The contribution of any item to the test reliability coefficient, or to the test validity coefficient, cannot be determined independently of the characteristics of all the other items in the test.

The quantity $P_i'^2/P_iQ_i$ in Equation (9) is the contribution of item i to the information function of the test. For this reason, it is called the *item information function.* These functions are additive. Figure 2 in Hambleton and Cook (1977) shows the item information functions for five verbal items. The information function for a test consisting of these items is obtained by adding up the ordinates of the five curves.

The following procedure for building a new test, first suggested by Birnbaum, is an excellent one. The procedure operates on a pool of items that have already been calibrated, so that we have the item information curve for each item.

(1) Decide on the shape desired for the test information function. (Remember that this information function is inversely proportional to the squared length of the asymptotic confidence interval for estimating ability from test score. See Hambleton & Cook.) What accuracy of ability estimation is required of the test at each ability level? The desired curve is the *target information curve.*
(2) Select items with item information curves that will fill the hard-to-fill areas under the target information curve.
(3) Cumulatively add up the item information curves, obtaining at all times the information curve for the part-test composed of items already selected.
(4) Continue (back-tracking if necessary) until the area under target information curve is filled up to a satisfactory approximation.

Figure 1 illustrates how a test might be built to approximate a hypothetical target information curve (heavy line). The item information curves for fifteen items are shown at the bottom. The three middle items are selected first. After this, items are added a few at a time, proceeding outwards from the middle. Part-test information curves are shown for 3-, 7-, and 11-item tests. The final information curve for the 15-item test approximates the target information curve.

REDESIGNING A TEST

A different procedure will be of value when the task is to improve or change an existing test, rather than building a completely new one. The procedure is best explained by citing a concrete example.

Recently, it was desired to change slightly the characteristics of the College Entrance Examination Board's Scholastic Aptitude Test, Verbal Section. It was desired to make the test somewhat more appropriate at low ability levels without impairing its effectiveness at high ability levels. The possibility of simultaneously shortening the test was also considered.

The first step was to estimate the item parameters for all items in a typical current form of the Verbal test. The second step was to compute information curves for variously modified hypothetical forms of the Verbal test. Each of these curves was compared to the information curve of the actual Verbal test. The ratio of the two curves is their *relative efficiency,* which varies as a function of ability level.

Consider a hypothetically modified test formed by adding to the actual test, five extra items having the same statistical properties as the five easiest items in the actual test. The relative efficiency of the modified test compared to the actual test is shown by curve 2 in Figure 2. As we might expect, the modified test measures a little better at low ability levels without losing or gaining anything at high ability levels.

Consider next the effect of eliminating a block of five items of medium difficulty

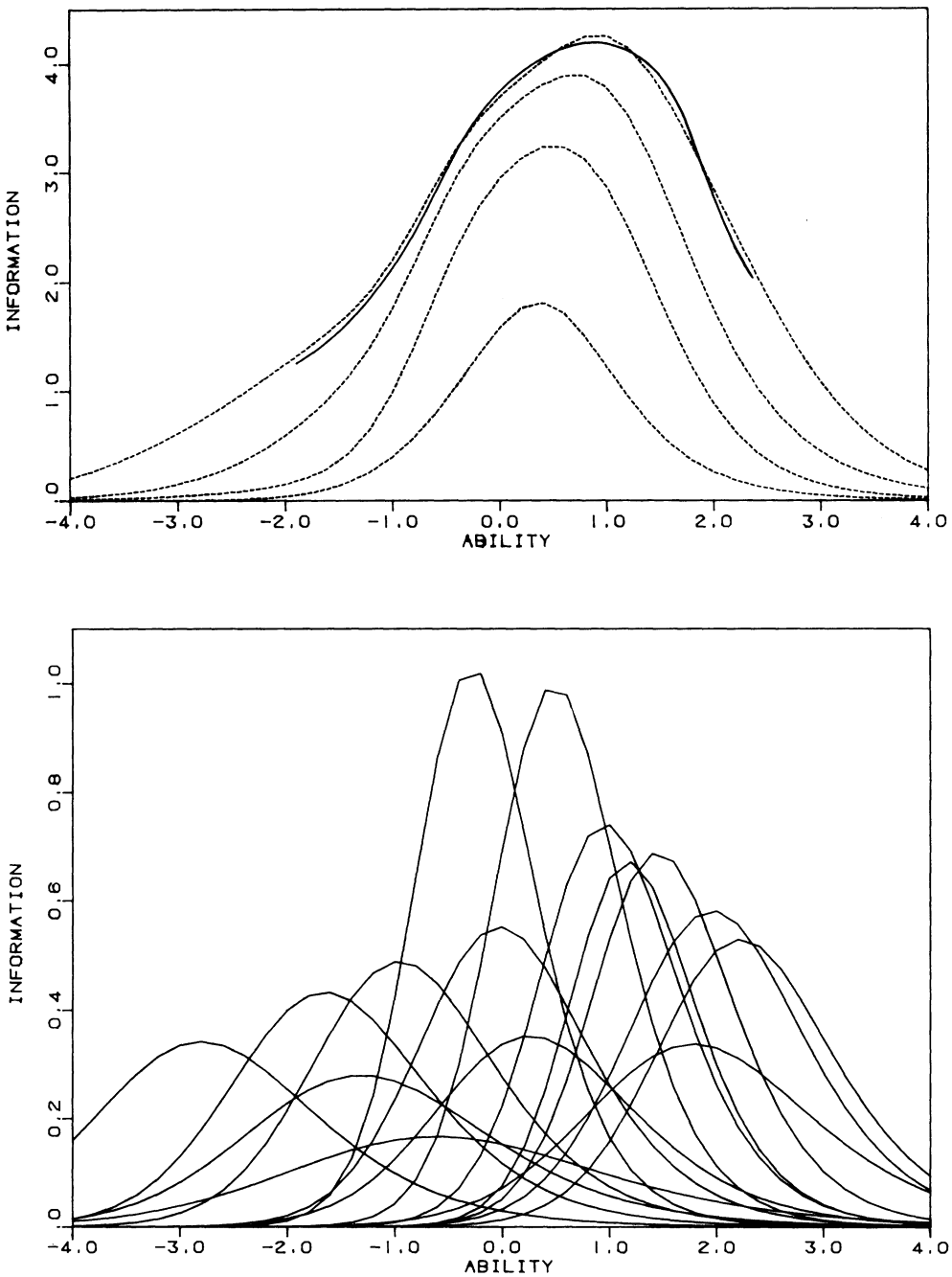

Figure 1 (top). Target information curve (solid) and subtest information curves (n = 3, 7, 11, 15). Figure 1 (bottom). Information curves for 15 items used to approximate target curve.

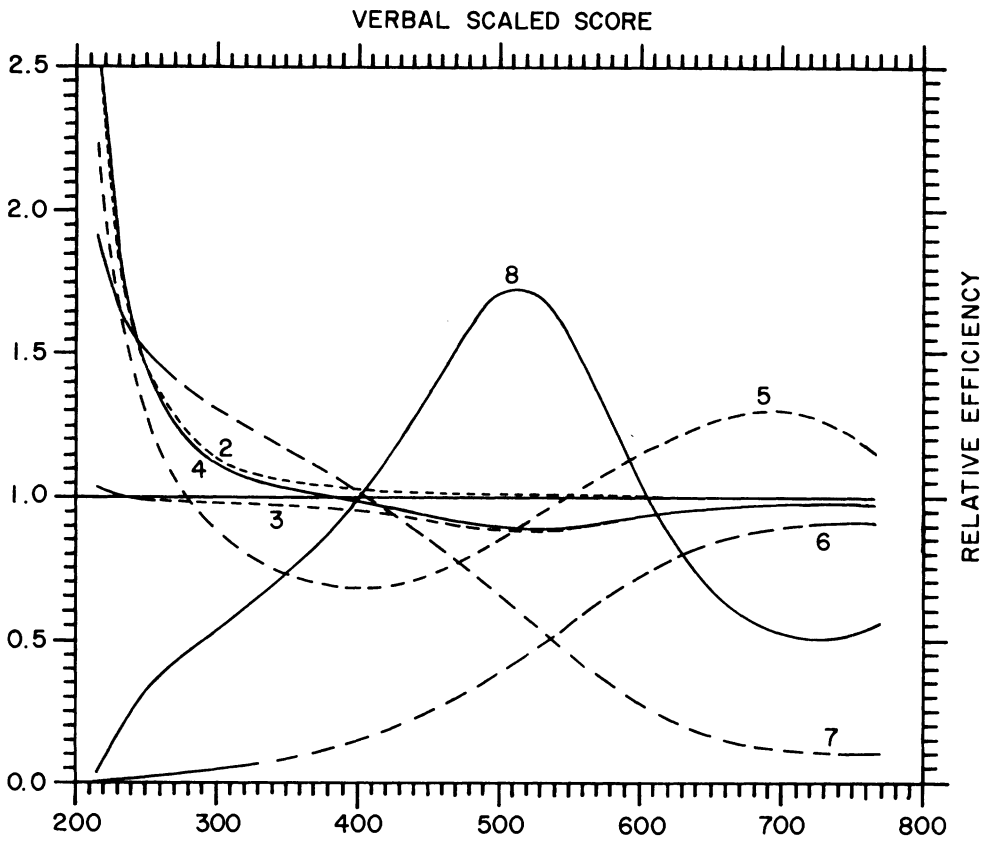

Figure 2. Relative efficiency of various modified SAT Verbal tests.

from the actual test. Measurement is impaired, mainly at middle ability levels, as shown by curve 3.

Now, suppose we eliminate the five medium-difficulty items at the same time adding the five easy items. The relative efficiency of the resulting test is shown by curve 4.

We can continue in this way, trying out as many hypothetically modified tests as we wish, until we have found one that comes close enough to the result desired. Curve 5 shows the effect of removing all reading items and then adding items parallel to the remainder to bring the test back to its original length. The remaining curves in Figure 2 serve to illustrate some basic principles of test design, which might not otherwise be evident.

Curve 6 shows the effect of throwing away the easiest half of the existing test. The measurement of high level examinees is not much affected, indicating that for them the discarded items are simply deadwood, a waste of testing time.

Curve 7 shows the effect of throwing away the hardest half of the existing test. In contrast to curve 6, we now find that the measurement of low ability examinees is much improved, indicating that the hard items are not merely deadwood but are definitely harmful for measurement of low-ability examinees. The reason is that such examinees guess at random on hard items, introducing a source of measurement error that impairs whatever measurement is achieved by the easier items.

Curve 8 shows the efficiency of a hypothetical 'peaked' test composed of items that differ from the actual test items only in difficulty, b_i. For curve 8, all the b_i are equal. The actual SAT Verbal test differs from the peaked test because the actual test contains many easy items (for the benefit of the low ability students) and many hard items (for the high ability students).

OPTIMAL SCORING WEIGHTS FOR ITEMS

If items are scored 0 or 1, as assumed throughout this paper unless otherwise specified, and if test score y is a weighted sum of item scores, then the score information function given by Equation (8) has the form given by Hambleton and Cook (1977, Equation (1)). If optimal weights (Hambleton & Cook, Equation (4)) are used, then the score information function is found to be identical with the maximal information function $I(\theta, \hat{\theta})$, given here as Equation (9). This means that an optimally weighted composite score is asymptotically as good a measure of ability as is the maximum likelihood estimator $\hat{\theta}$.

The optimal item-scoring weight is a function of examinee ability. The functions $w_i(\theta)$ for the five items of Hambleton's Figure 2 are shown in Figure 3 (reproduced from Lord, 1968, by permission of *Educational and Psychological Measurement*). The main point to note here is that the items that should be heavily weighted for measuring high-ability examinees are not necessarily the items that should be heavily weighted for measuring low-ability examinees.

For measuring high-ability examinees under the three-parameter logistic model, the optimal scoring weight for most items (for items that are easy for such examinees) is proportional to the item discriminating power a_i. When measuring low-ability examinees, difficult items should receive near-zero scoring weight, regardless of their a_i parameter (see the discussion of curve 7, Figure 2 in the preceding section).

Item-scoring weights that are optimal for a particular examinee can never be determined exactly, since we do not know the examinee's ability θ exactly. For the logistic model the optimal weight is

$$w_i(\theta) = \frac{Da_i}{1 - c_i} \frac{P_i(\theta) - c_i}{P_i(\theta)}. \tag{10}$$

A crude procedure for obtaining item-scoring weights is to substitute the conventional item difficulty p_i (proportion of correct answers in the total group of examinees) for $P_i(\theta)$ in Equation (10). A crude procedure would use the resulting weight for scoring item i on all answer sheets regardless of examinee ability level. Since D = 1.7 is a constant, we can drop it and use the weight

$$w_i = \frac{a_i}{1 - c_i} \frac{p_i - c_i}{p_i}. \tag{11}$$

This same item-scoring weight, except for the a_i, was recommended on other grounds by Chernoff (see Lord & Novick, 1968, p. 310).

If there is no guessing, $c_i = 0$ and the item scoring weight is equal to a_i, the item discriminating power. If there is guessing, Equation (11) gives low scoring weight to difficult items. The justification is that a correct response to a difficult item may indicate lucky guessing rather than high ability.

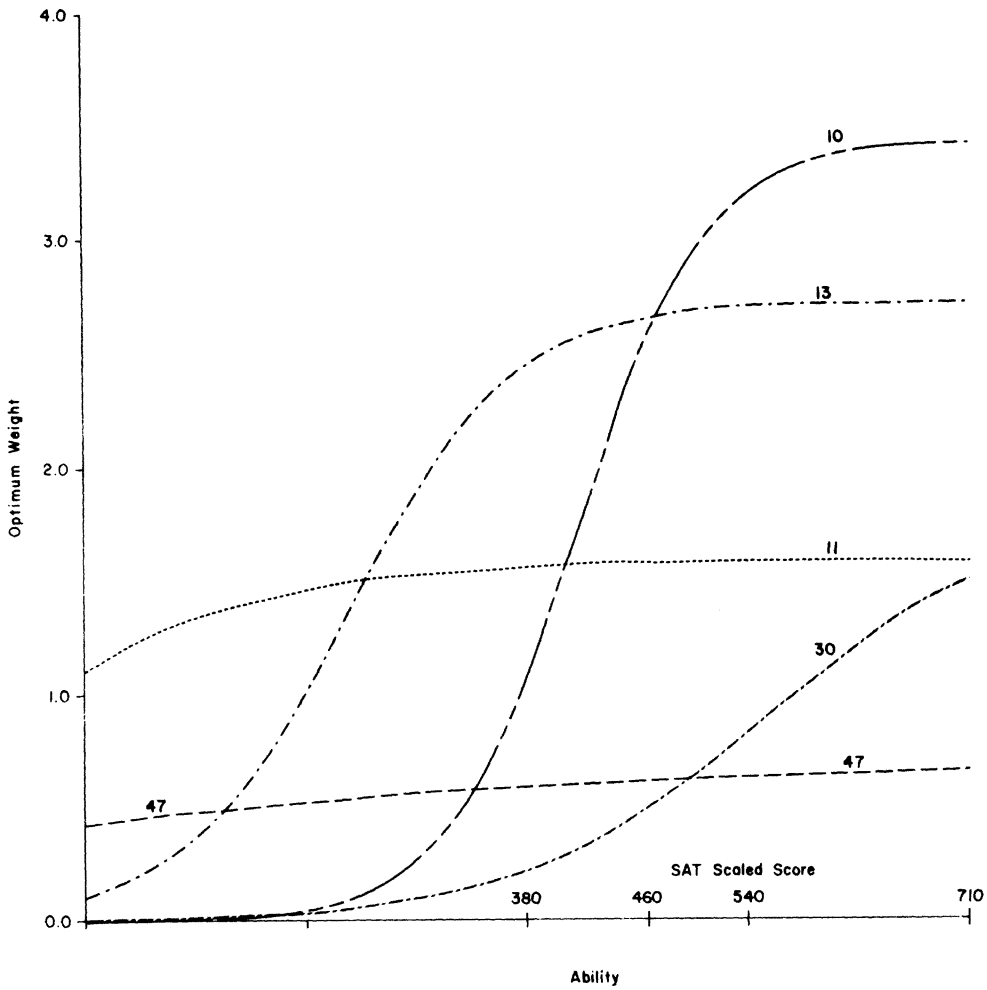

Figure 3. Optimum item weight as a function of ability level where measurement is desired, as estimated for five SAT Verbal items.

A better procedure for determining scoring weights for a conventional test might be somewhat as follows.

1. Score the test in the usual way.
2. Divide the examinees into three groups according to the usual scores.
3. Separately for each subgroup, use Equation (11) to find a roughly optimal scoring weight for each item.
4. Rescore all answer sheets using different item-scoring weights (from step 3) in each subgroup.
5. Equate the three score scales obtained from the three sets of scoring weights. Conventional equating methods may be used for this.
6. Use the equating to place everyone on the same score scale.

The foregoing procedure should improve measurement effectiveness, since each answer sheet is scored with weights roughly appropriate for the examinee's ability

level. Too much should not be expected, however. If only a third of the items in a test are useful for measuring examinees at a certain ability level, no amount of statistical manipulation will make the test a really good one for such examinees.

MASTERY TESTING

A primary purpose of mastery testing is to determine whether or not each examinee has reached a certain required level of achievement, denoted here by θ_0. The optimal item-scoring weight for determining mastery is $w_i(\theta_0)$,where $w_i(\theta)$ is the optimal weight discussed in the preceding section. In mastery testing, since $w_i(\theta_0)$ does not depend on examinee ability, the same scoring weight can be used for all examinees. The weights $w_i(\theta_0)$ are optimal for all examinees near θ_0 in ability.

The examinee's weighted composite score $\sum_i w_i(\theta_0)u_i$—where u_i denotes raw score on item i—is to be compared to a predetermined cutting score. If the examinee's score is higher, the procedure labels him as a master; otherwise, as a nonmaster. The cutting score is prechosen so as to limit to some predetermined level, the probability that a nonmaster will be erroneously labelled as a master.

The reader is referred for further details to Birnbaum (1968, Chapt. 19). Birnbaum tells how to select the best items for building a mastery test, how to determine the number of items needed to reach a decision, how to proceed when there is more than one category of mastery to be determined, and how to evaluate the effectiveness of mastery tests. This seems to be a particularly felicitous application of item response theory, since mastery tests are generally very nearly unidimensional.

TAILORED TESTING

If the group to be tested is sufficiently heterogeneous, it is impossible for a conventional test to measure accurately at both high and low ability levels at the same time. This is apparent from the preceding sections on redesigning a test and on optimal item-scoring weights. To obtain effective measurement at low (high) ability levels, we need easy (hard) items. When number-right scores or conventional formula scores are used, the hard items not only waste the time of the low-ability examinees, they impair whatever measurement of these examinees would otherwise be effected by the easy items. The use of item-scoring weights can eliminate the damage done by the hard items, but cannot produce the effective measurement that would have been achieved by a full-length test composed entirely of easy items.

If we wish to measure accurately throughout a wide range of ability, we need to match the difficulty level of the items administered to the ability level of the examinee tested. This is individualized or *tailored testing*. Each different examinee usually takes a different set of items.

Effective tailored testing involves two repeated basic steps (or some approximation to them). Neither of these steps could be effected by classical test theory. Independently for each examinee being tested:

1. Before the next item is selected and administered, estimate the examinee's ability from hir responses up to this point.
2. Among all available items not yet administered, select the item likely to measure most effectively at the examinee's estimated ability level. In effect, this means pre-

dicting the examinee's chance of success on each item in the pool at each step of the testing.

After the administration is complete, all examinees must be placed on the same score scale. This is accomplished by using the final ability estimates obtained from step 1 above.

Tailored tests can be designed to cover as wide a range of ability as desired. This is illustrated by the *Broad-Range Tailored Test of Verbal Ability* (Lord, 1976), which is designed to span the entire range from typical fourth-grade pupils through top-level graduate school students, placing all examinees on the same score scale. The items for this test have been selected and calibrated, the tailoring procedure has been specified, administration of the test has been simulated on the computer, and the results have been evaluated. The test is quite effective for simulated examinees. Actual testing of actual people, however, is not at present feasible with this test because the necessary computer programs have not been written.

There is much more to be said about the application of item response theory to tailored testing. The reader should refer to Urry (1977) for detailed discussion and for a description of the actual computer implementation and administration of a tailored test, and for practical evaluation. Other recent work is summarized in C. L. Clark (1976), in Weiss (1976), and in Lord (1974a). Also see Linn, Rock, and Cleary (1972), Mussio (1973), McBride (1976), and Cudeck, Cliff, Reynolds, and McCormick (1976).

TWO-STAGE TESTING

Another way of matching the difficulty of the items administered to the ability of the examinee is by using two-stage testing. The first stage is used to route the examinee to one of several second-stage tests. The difficulty of the second test is chosen to match the examinee's level of performance on the routing test.

Two-stage testing can be implemented without a computer. However, the routing test must be scored before the second test can be administered.

What is the relation of two-stage testing to item response theory? In principle, a two-stage test can be thrown together, administered, and scored without using item response theory. Even if thoughtfully planned, however, a test designed in this way is likely to prove less effective than a conventional test, in the writer's experience.

Design decisions that must be made by the psychometrician include the following:

1. length of routing test,
2. length of second-stage test(s),
3. number of second-stage tests,
4. difficulty level of the routing test,
5. difficulty level of each second-stage test,
6. method of scoring routing test,
7. cutting scores on routing test for each second stage test,
8. method of scoring second-stage test,
9. method of combining scores from first and second stages.

If the routing test is too short, routing is inefficient; if it is too long, the second-stage test will be shorter than it should be for a given amount of total testing time. If there are too few second-stage tests, tailoring will be inadequate; if there are many second-stage tests, measurement will not suffer, but test development costs will be unnecessarily large.

The choice of difficulty level for each second-stage test and the choice of the corresponding cutting scores on the routing test are most difficult matters. An unfortunate choice of cutting scores may lead to a situation where most examinees are routed to the same second-stage test (Betz & Weiss, 1973), or to other difficulties. Even using item response theory, it seems necessary to choose an overall design only after simulating a variety of designs, evaluating each, and repeating this procedure in an enlightened trial and error process.

Such use of item response theory is illustrated in Figure 4 (reproduced by permission of *Psychometrika* from Lord, 1971). The broken lines show the information functions for three different two-stage procedures. The solid curves are for comparison purposes: 'Up-and-Down' is a tailored-testing procedure, 'Standard' is a conventional test with all items at the same difficulty level.

All information functions are obtained by formula, not by Monte Carlo methods. This makes it possible to try out and tentatively evaluate hundreds of two-stage test designs on a computer, much more easily than a single two-stage test could be evalu-

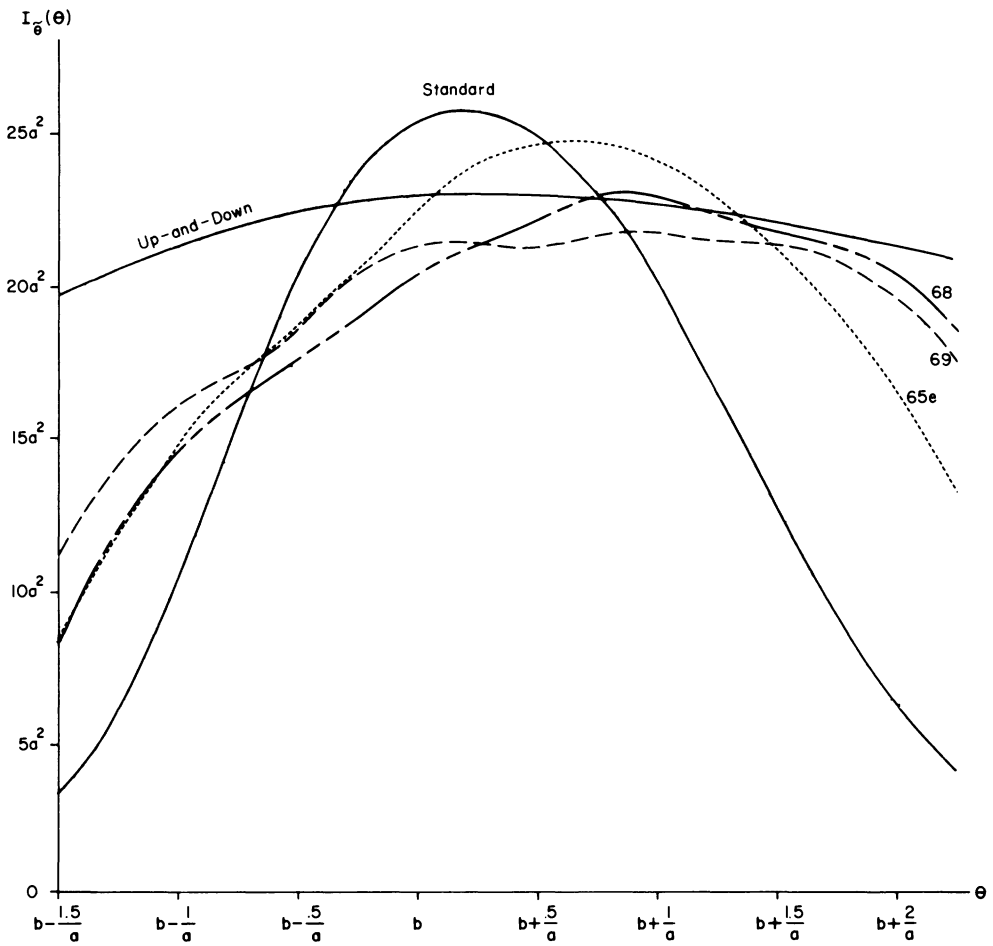

Figure 4. Two-stage procedure, n = 60, c = .2.

ated by administration in the field to real examinees. Only the designs with the best information curves need be tried out in actual testing.

Formulas, discussion, and other examples are given in Lord (1971). Larkin and Weiss (1975) discuss an experimental application of two-stage testing and cite some earlier literature.

Multilevel Testing

Two-stage testing is more convenient when the routing test can be given in advance of the regular testing as a take-home self-scored test. In this case, the score on the routing test is used to route the examinee, but does not otherwise affect hir final score.

A still more convenient version of two-stage or multistage testing provides examinees with a test composed of several separate levels and instructs them to route themselves—to choose among the levels subject to certain specified restrictions. This kind of testing may be exceptionally efficient, since it takes advantage of whatever knowledge examinees may have about their own ability levels.

A practical example of multilevel testing is discussed in this journal issue by Marco (1977). For this reason, it is not discussed further here.

EQUATING

When two tests are approximately parallel, the equation relating equivalent scores on the two tests is approximately linear, so that simple and convenient, conventional linear equating methods can be used. The discussion here will deal with nonparallel tests, unless otherwise specified. If we do not know that two tests are really parallel, the safe procedure is to treat them as nonparallel and to use nonlinear methods.

Scores y on test Y are equated to scores x on test X by a transformation $y^* \equiv y^*(y)$ that transforms 'raw' y scores 'on to the x scale.' Strictly speaking,

Definition. *Transformed scores* y^* *and raw scores* x *can be called 'equated' if and only if it is a matter of indifference to each examinee whether he is to take test X or test Y.*

Equatings satisfying this definition cannot always be found. This definition seems nevertheless to be definitely required by the ways in which supposedly equated scores are actually used in educational practice.

Under this definition, tests measuring different traits or abilities cannot be equated. Neither can raw scores on unequally reliable tests of the same trait or ability. The last statement must be true, since otherwise there would be no reason to build reliable tests—an unreliable test could be made to yield scores interchangeable with scores from a reliable test simply by equating. It is not a matter of indifference to high-level examinees whether they take a reliable or an unreliable test. Their abilities will be made evident by a reliable test, but may not be made evident by an unreliable one.

If tests X and Y are of different difficulty, the relation between their true scores is necessarily nonlinear, because of floor and ceiling effects (also see discussion of Equations (12) and (13) below). If two tests have a nonlinear relation, it is implausible that they should be equally reliable for all subgroups of examinees. This leads to the awkward conclusion that, strictly speaking, *observed* scores on tests of different difficulty cannot be equated. Faced with this conclusion, we need to think clearly in what follows.

A necessary but *not* a sufficient condition for y^* and x to be 'equivalent' scores (i.e., to be equated) is that y^* and x have identical frequency distributions in all groups and

subgroups of interest. In this paper, the term equating will sometimes be used loosely to describe a situation where the equating is only approximate.

Many different designs may be used in an equating study. Item response theory can be used in somewhat different ways, depending on the design. A design based on two nonequivalent groups of examinees will be used here to illustrate what can be done. This is of special interest since conventional equating methods are not strictly appropriate when nonparallel tests having a nonlinear relationship are administered to nonequivalent groups.

In such situations an 'anchor' test must be used to measure and take into account the difference in ability between the two groups. Each group takes one of the tests to be equated, followed by the anchor test, which must be a measure of the same ability as the other two.

The illustrative example carries through the equating of two calculus tests: 1) a 45-item achievement test from the regular College Board advanced placement program (AP) and 2) a somewhat easier 50-item test from the College Level Examination Program (CLEP). The anchor test is provided by 17 items that appear in both AP and CLEP. The equating study was proposed and planned by Dr. Gary Marco.

The equating utilizes a sample of 1260 examinees who took AP Calculus in May 1972 and an entirely separate sample of 1209 examinees who took CLEP Calculus in May 1974. Of these, 243 (130 from AP, 113 from CLEP) who failed to answer more than 26 items were omitted from the calculations (this was a purely arbitrary decision based on the grounds that such examinees may not be adequately measured). The two samples differed noticeably in ability level. The anchor test measures and corrects for this difference; also for the effects of omitting the 243 examinees.

LOGIST[1], a computer program for estimating item and ability parameters, is briefly described in Hambleton and Cook (1977). A LOGIST computer run was made simultaneously on all 1260 + 1209 − 243 = 2226 examinees and on all 45 + 50 − 17 = 78 items, obtaining one estimated ability parameter $\hat{\theta}$ for each examinee and three estimated item parameters $\hat{a}$, $\hat{b}$, and $\hat{c}$ for each item.

The $\hat{\theta}$ values are not used in the true-score equating procedure that follows. This is appropriate, since such an equating is supposed to be independent of the group tested.

True-Score Equating

There is an exact mathematical relationship between ability θ and number-right true score ξ on test X:

$$\xi \equiv \xi(\theta) \equiv \sum_{i=1}^{n_x} P_i(\theta), \tag{12}$$

where n_x is the number of items in test X. This is really just a repetition of Equation (2). Similarly, for number-right true score η on test Y,

$$\eta \equiv \eta(\theta) \equiv \sum_{j=1}^{n_y} P_j(\theta). \tag{13}$$

[1]LOGIST is available from the writer. We will place it on your magnetic tape if you supply the tape and necessary tape specifications.

These are exact mathematical, not statistical, relations. Equations (12) and (13) are parametric equations for the relation between true scores for tests X and Y; in other words, if we substitute any value for θ in Equations (12) and (13), they produce equated values of true scores η and ξ.

Alternatively, we can eliminate θ from Equations (12) and (13), obtaining an exact mathematical equation relating ξ and η. Since this is an exact relation, it certainly accomplishes the equating of the true scores on tests X and Y. The relationship cannot be written in simple algebraic form, but equated values can easily be computed numerically if the $P_i(\theta)$ and $P_j(\theta)$ are known. Note that the relation between true scores is necessarily nonlinear unless tests X and Y are strictly parallel, item by item.

In practice, we do not know $P_i(\theta)$ and $P_j(\theta)$, but we do have good estimates $\hat{P}_i(\theta)$ and $\hat{P}_j(\theta)$, obtained for each item by substituting â, b̂, and ĉ for the unknown parameters a, b, and c. The computer can easily compute the (estimated) *true-score equating* of test X and test Y in this way.

The dashed line in Figure 5 shows the true-score equating estimated for AP and CLEP. The other curves in this figure are discussed in a later section.

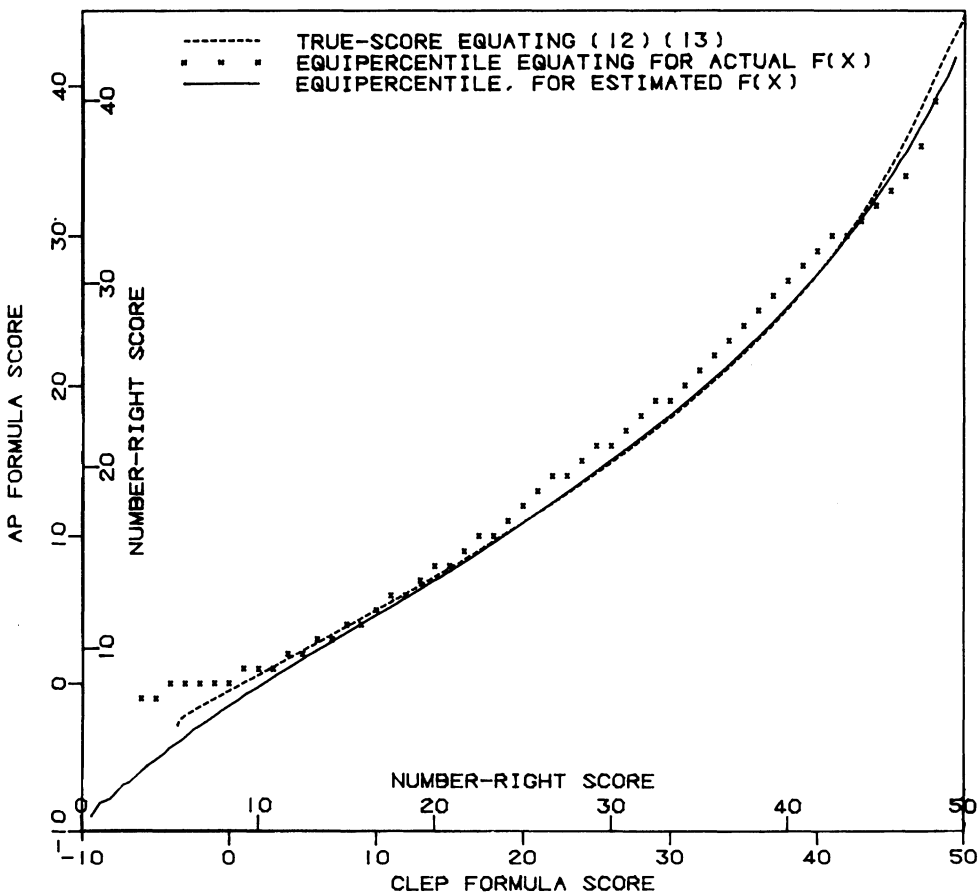

Figure 5. Lines of relationship for AP and CLEP.

Observed-Score 'Equating'

If the assumptions of item-response theory hold, the preceding section surely accomplishes the desired true-score equating. The problem is that in practice there is no really correct way to use the results obtained, since we never have any examinee's true score. We cannot correct this by substituting estimated true scores for their true values, since the relation between estimated true scores will in general be different from the relation between their true values. The estimated true score itself is just a new kind of observed score.

If the difference between true-score and observed-score equating seems unimportant, consider the following. In practical applications, we need to equate number-right scores for the entire range from 0 to n. However, the possible range for true scores is only from $\xi = \sum_i c_i$ (the pseudochance level) to $\xi = n$. Thus, true-score equating can never produce a line of relationship extending below the limits set by the item pseudochance levels. For number-right scores, the lower limit is 6.7 on AP, 7.2 on CLEP; for formula scores the limits are respectively -2.9 and -3.5 (see the section below on formula scoring).

If we know the distribution $g(\theta)$ of θ in the combined group tested, we can use the estimated item parameters of each test to estimate by means of Equations (1) and (4), the frequency distribution of the number-right observed scores on that test that would be found in the combined group, if everyone had taken that test. This can be done in spite of the fact that no examinee took both AP and CLEP.

The observed scores can then be approximately equated by an ordinary equipercentile equating procedure applied to their estimated frequency distributions in the combined group. The equating here is called approximate because the result will not be independent of the distribution of θ in the group tested, and thus will not satisfy a basic principle of equating.

The actually obtained distribution of $\hat{\theta}$ can be used as an approximation to the $g(\theta)$ needed for the procedure just described. However, the distribution of estimated θ is not really the same as a good estimated distribution of θ. The method of Lord (1969; 1974b, Figure 2) will usually provide a good estimate of $g(\theta)$, provided there are not too many omitted responses on the answer sheets.

The solid curve in Figure 5 was obtained by the method just outlined using the distribution of $\hat{\theta}$. It agrees very well with the true-score line of relation found in the preceding section. The observed-score line of relation extends well below the chance level.

The curve of crosses in Figure 5 was obtained by conventional methods. The AP was equated to the 17-item (internal) anchor test by the equipercentile method; CLEP likewise. Scores equated to the same anchor-test score were then assumed to be equated to each other. This last step would be a valid one if the equipercentile equating of raw scores remained the same from group to group. Since strictly speaking it does not, the line of crosses in Figure 5 is not a strictly correct equating.

The discrepancy in Figure 5 between the other curves and the conventional equating may also be due to the following facts: 1) As explained earlier, the item-response-theory equatings omitted 243 examinees who were included in the equipercentile equating. This is unfortunate for purposes of the present comparison; however, a proper equating is supposed to be invariant from group to group. 2) The item-response-theory equatings are based on a prediction of what would happen if there were no omitted items (see section below on omitted responses).

In theory, item response methods are capable of estimating the equipercentile line of relation between raw scores when the two tests to be equated are not parallel, are given to nonequivalent groups, and everyone takes an anchor test. Strictly speaking, no other method known to the writer can accomplish this. The resulting line of relation depends on the groups tested, however.

Additional Illustration

The foregoing example was chosen because it deals with a practical equating problem. For just this reason, however, there is no satisfactory way to check on the accuracy of the results obtained. The results obtained by conventional methods cannot be justified as a criterion.

The following example was set up so as to have a proper criterion for the equating results. Here, unknown to the computer procedure, test X and test Y are actually the same test. Thus we know in advance what the line of relation should be.

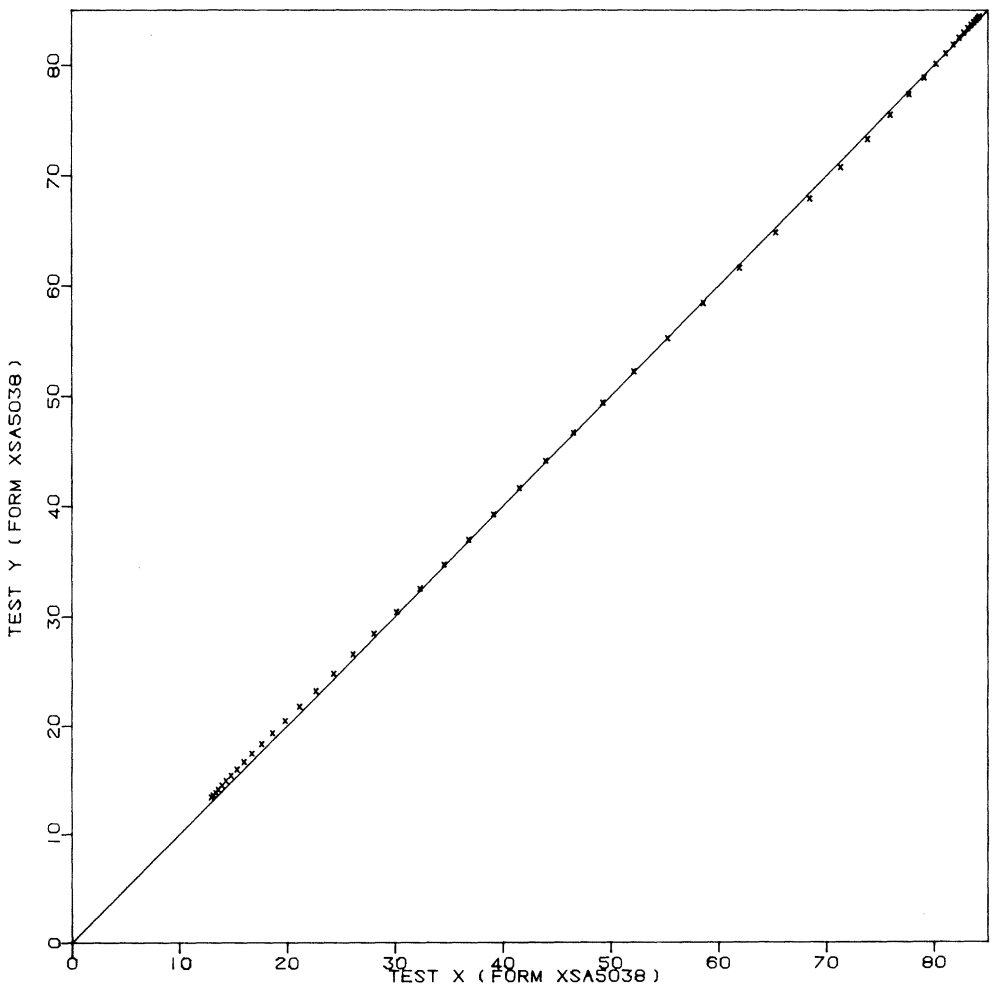

Figure 6. Estimated equating (crosses) between 'Test X' and 'Test Y', which actually are identical.

Test X is the 85-item verbal section of the College Board SAT, Form XSA, administered to a group of 2802 college applicants in a regular SAT administration. Test Y is the same test administered to a second group of 2763 applicants. Both groups also took a 39-item verbal test mostly (but not entirely) similar to the regular 85-item test. The 39-item test is used here as an anchor test for the equating.

The two groups differed in ability level. The proportion of correct answers given to typical items is lower by roughly 0.10 in the first group than in the second.

The equating was carried out exactly as described in the section titled *True-Score Equating*. One LOGIST run was made simultaneously for all 85 + 85 + 39 = 209 items and for all 5565 examinees. The resulting line of relationship between true-scores on test X and test Y is shown by the crosses in Figure 6. It agrees very well with the 45-degree line, also shown, that should be found when a test is equated to itself.

DISCUSSION

We have here a well-known dilemma. Statisticians are often accused of providing answers to the wrong questions. They answer some question that can be answered, instead of the really important question asked, which cannot.

If two tests are not strictly parallel to start with, it will ordinarily be impossible, by definition, to equate their observed scores. No single 'equating' can be equitable within all subgroups of examinees. Since a uniformly equitable result does not exist, what shall we do instead?

Either the exact true-score equating can be used with observed scores, or else an inexact observed-score equating can be used. The real problem is that we have no criterion for choosing.

Equating is not a prediction problem. If it were, we would be using regression methods. These are not appropriate here.

At present, we do not have enough experimental results like those in Figures 4 and 5 to enable us to generalize about what may be expected from the methods under discussion. A large-scale equating study sponsored by the College Entrance Examination Board and by Educational Testing Service is currently in progress, under the direction of G. Marco and E. Stewart, comparing a variety of different equating methods for very many different sets of data. Perhaps this experience will lead to some criterion for choosing between the suggested procedures.

Omitted Responses

LOGIST is designed to operate effectively and appropriately on answer sheets containing both omitted and not-reached items. The method used to do this is discussed in Lord (1974b). The ability and item parameter estimates obtained by LOGIST are supposed to be about the same, whether the examinee omits certain items or guesses at random and whether or not the examinee has time to finish the test. All equations in this paper that involve summations or products over $i = 1, 2, \ldots, n$ give the result that would be expected if the examinee(s) were to answer all n items.

In many cases these summations or products could instead easily be taken only over those items to which the examinee actually responded. The results would then apply to real-world situations.

This approach cannot be applied to the first illustration of equating given above. The reason is that, although we can predict the number-right score of an examinee on a test

that (s)he has not taken assuming (s)he answers all the test items, we cannot predict hir omitting and not-reaching behavior on such a test. This is not as serious for equating purposes as it might at first appear, since an equating is supposed to be independent of the group of examinees. If an equating is correct for examinees who answer all test items but is not correct for examinees who do not, then no single equating can be valid for use with this test in any case.

FORMULA SCORING

A *formula score* z is a number-right score (x) minus a fraction $\left(\frac{1}{A}\right)$ of the number wrong (w):

$$z \equiv x - w/A. \tag{14}$$

The College Board scaled score, for example, is a linear transformation of a formula score.

The first thing to be noted here is that when there are no omitted responses, formula scores are perfectly correlated with number-right scores. If everyone answers all n items, $w = n - x$ and

$$z \equiv (1 + 1/A)x - n/A.$$

Thus, if there are no omitted responses, any result obtained for number-right scores can be quickly translated to apply to formula scores. As explained in the preceding section, many or most applications of item response theory represent what would be expected if everyone answered all the items in the test. For this situation, it is as easy to obtain results for formula scores as for number-right scores. This explains why the base line of Figure 5 is scaled both for number-right and for formula scores.

There is another way in which item response theory can be of use for evaluating formula-scored tests. Under certain plausible assumptions, we can find the relative efficiency (see Hambleton & Cook, 1977, p. 87) of a formula score compared to that of a number-right score. It is a function of the pattern of omits and of the P_i (Lord, 1975):

$$\text{R.E.}\{z, x\} = \sum_{i=1}^{n} P_iQ_i \Big/ \left[\sum_{i=1}^{n} P_iQ_i - \frac{A-1}{A^2}\, \mathcal{E}(n_o \mid \theta)\right]. \tag{15}$$

Here A is the number of alternative responses per item and $\mathcal{E}(n_o \mid \theta)$ is the regression of n_o, the number of omitted items, on θ. This regression is approximated by finding the (usually nonlinear) regression of n_o on $\hat{\theta}$ for the sample of examinees at hand.

In Equation (15), the numerator and denominator are the same except for the subtracted term $(A - 1)\mathcal{E}(n_o \mid \theta)/A^2$. Since this term cannot be negative, the relative efficiency of formula scores cannot be less than 1.0. The relative efficiency of formula scoring increases as the number of omits increases, also as the number of choices per item decreases. The number of omits tends to be near zero for high ability students, greater than this for low ability students. Thus formula scoring is less important for measuring high-ability than low-ability students.

DO TWO TESTS MEASURE THE SAME TRAIT?

Equation (1) shows how we can generate $f(x \mid \theta)$, the distribution of number-right score for fixed ability level θ. If we have two different tests X and X′ that measure the same trait, a similar equation holds for $f'(x' \mid \theta)$. For fixed θ these two distributions are independent, so we can compute the joint distribution for fixed θ

$$f(x, x' \mid \theta) = f(x \mid \theta) f'(x' \mid \theta). \tag{16}$$

Then, as in Equation (4), we can approximate the joint distribution of x and x′ for an entire group of examinees of varying ability by computing

$$\hat{f}(x, x') = \frac{1}{N} \sum_{a=1}^{N} f(x, x' \mid \hat{\theta}_a), \tag{17}$$

(or better, by the method discussed in Lord, 1969). This joint distribution can be computed even if no one has taken both tests X and X′.

If a group of examinees has taken both tests, the joint distribution predicted by Equation (17) can be compared with the actually-observed scatterplot. If one is confident from this and other empirical checks that the item-response model holds for tests X and X′ separately, then a significant discrepancy between observed and predicted $f(x, x')$ can be taken as evidence that the two tests are not measures of the same trait. Such conclusions are important in studies of construct validity and elsewhere.

If tests X and X′ are of different difficulty, it is likely that the relation between x and x′ is nonlinear. In this case the usual correlation corrected for attenuation is inappropriate and the method of Equation (17) may be preferable. A low corrected correlation could be due to a curvilinear relationship, rather than to any real difference between the abilities measured by the test.

MEASURING THE APPROPRIATENESS OF MULTIPLE-CHOICE TESTS

Certain individuals may not be properly measured by a particular test. This can occur because these individuals have language difficulties, or have an unusual background in some other way, or because they do not understand the test directions, or perhaps even because of cheating. It would be valuable to be able to locate such individuals and separate them from the great majority whom the test measures appropriately.

For any individual, the asymptotic standard error of hir $\hat{\theta}_a$ can usually be determined approximately by LOGIST or other computer programs. If this standard error is too large, the test will not measure this individual accurately; perhaps hir test results should be disregarded.

If the test is inappropriate for the examinee for any of the reasons mentioned, it is likely that hir pattern of responses will be atypical and will not be well fitted by the same model that fits the other examinees' responses. Thus any measure of the fit of the model to the examinee's answer sheet may pick out examinees for whom the test is inappropriate.

Michael Levine has pointed all this out. With Donald Rubin, he has tried out various indices for picking out examinees for whom the test is inappropriate. The reader is referred to Levine and Rubin (1976) and to forthcoming reports of their work.

ITEM BIAS

Each item response curve in Figures 7, 8, and 9 represents the probability of success on a particular item as a function of ability level. The three figures are for three different items from the Verbal Scholastic Aptitude test. The solid curve in each figure is for a group of white students. The dotted curve is for a group of black students.

In Figure 7 we see that the high ability white students do better on this item than high ability black students, but that low ability black students do better on the item than low ability white students.

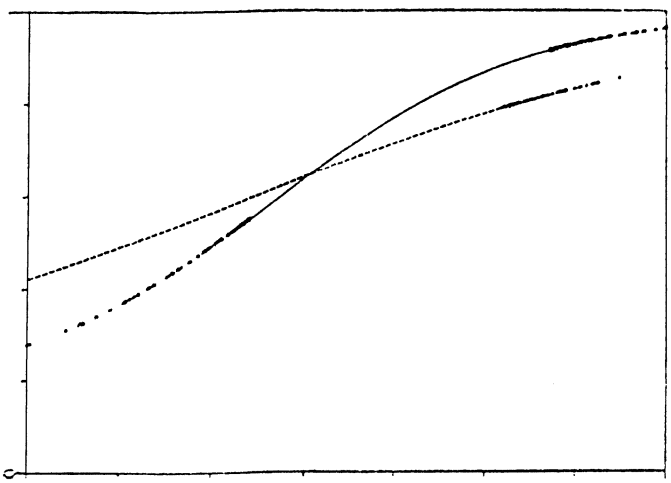

Figure 7. Black (dashed) and white (solid) item response curves for item 71.

Figure 8 shows a similar situation except that in this case the item is totally undiscriminating for black students. High ability black students (as determined by other items on the test) do no better on this item than do low ability black students.

Figure 9 shows a difficult item on which blacks do better than whites at every ability

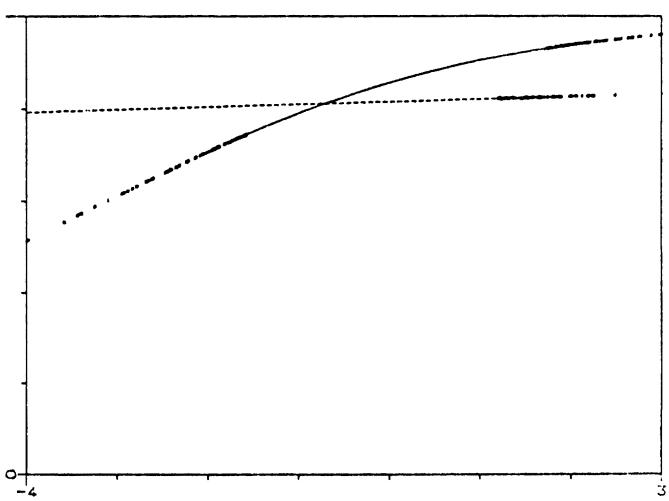

Figure 8. Item response curves for item 2.

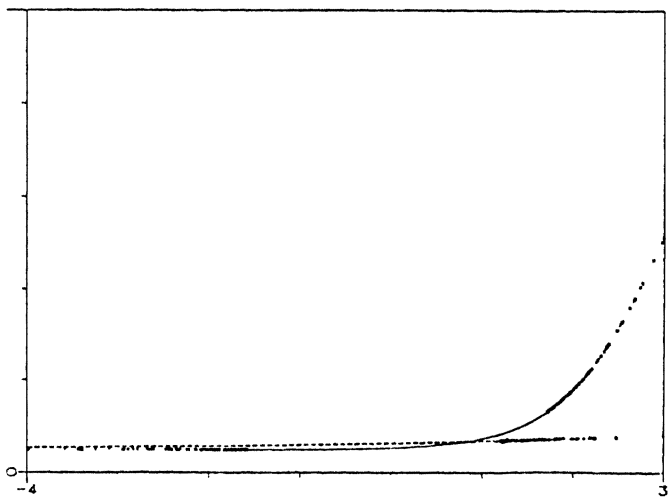

Figure 9. Item response curves for item 24.

level where there is a difference. There are, of course, other items on which whites do as well or better than blacks at each ability level.

Such items contain a bias, a somewhat complicated kind of bias. It would seem desirable to exclude such items from our tests as far as possible.

Let me emphasize that the curves shown here were picked simply because they did show a definite difference between black groups and white groups. Most of the items in the Verbal SAT do not show large biases of this kind.

These curves have only recently become available as a result of a study designed by Gary Marco. There has not yet been time to study the test items and compare them with the statistical results. A more complete report will be forthcoming. It is to be hoped that as a result of such studies, we will learn how to design items that do not show these kinds of bias.

REFERENCES

BETZ, N. E. & WEISS, D. J. An empirical study of computer-administered two-stage ability testing. Research Report 73-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, Minn., October 1973.

BIRNBAUM, A. Classification by ability levels. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968. pp. 436–452.

CLARK, C. L. (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing.* Washington, D.C.: United States Civil Service Commission, 1976.

CUDECK, R. A., CLIFF, N., REYNOLDS, T. J., & McCORMICK, D. J. Monte Carlo results from a computer program for tailored testing. Technical Report No. 2, Department of Psychology, University of Southern California, Los Angeles, Calif., February 1976.

HAMBLETON, R. J. & COOK, L. *Journal of Educational Measurement*, 1977, **14,** 117–138.

LARKIN, K. C. & WEISS, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing. Research Report 75-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, Minn., February 1975.

LEVINE, M. V. & RUBIN, D. B. Measuring the appropriateness of multiple choice tests. Research Bulletin 76-00, Educational Testing Service, Princeton, N.J., 1976, in preparation.

LINN, R. L., ROCK, D. A., & CLEARY, T. A. Sequential testing for dichotomous decisions. *Educational and Psychological Measurement,* 1972, **32,** 85–95.

LORD, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement,* 1968, **28,** 989–1020.

LORD, F. M. Estimating true-score distributions in psychological testing (An empirical Bayes estimation problem). *Psychometrika,* 1969, **34,** 259–299.

LORD, F. M. A theoretical study of two-stage testing. *Psychometrika,* 1971, **36,** 227–242.

LORD, F. M. Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, and P. Suppes (Eds.), *Contemporary developments in mathematical psychology,* Vol. II. San Francisco: Freeman, 1974. pp. 106–126. (a)

LORD, F. M. Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika,* 1974, **39,** 247–264. (b)

LORD, F. M. Relative efficiency of number-right and formula scores. *British Journal of Mathematical and Statistical Psychology,* 1975, **28,** 46–50.

LORD, F. M. A broad-range tailored test of verbal ability. In C. L. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing.* Washington, D.C.: United States Civil Service Commission, 1976, pp. 75–78.

LORD, F. M. & NOVICK, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

MARCO, G. L. Unpublished report. Educational Testing Service, Princeton, New Jersey, 1977.

MC BRIDE, J. R. Bandwidth, fidelity, and adaptive tests. In *CATC-2 1975.* Atlanta, Ga.: Atlanta Public Schools, 1976. pp. 81–98.

MUSSIO, J. J. A modification to Lord's model for tailored tests. Unpublished doctoral dissertation, University of Toronto, 1973.

URRY, V. W. Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement,* 1977, **14,** 181–196.

WEISS, D. J. Computerized ability testing, 1972–1975. Final Report, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, Minn., April 1976.

AUTHOR

LORD, FREDERIC M. *Address:* Educational Testing Service, Princeton, NJ 08540. *Title:* Distinguished Senior Research Scientist and Chairman, Psychometric Research Group. *Degrees:* B.A. Dartmouth College, M.A. University of Minnesota, Ph.D. Princeton University. *Specialization:* Psychometric Theory and Methods; Mathematical and Applied Statistics.