DOCUMENT RESUME

ED 079 356 TM 002 955

AUTHOR	Lord, Frederic M.
TITLE	The Relative Efficiency of Two Tests as a Function of Ability Level.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	National Science Foundation, Washington, D.C.
REPORT NO	ETS-RB-73-41
PUB DATE	Jun 73
NOTE	16p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Standard Error of Measurement; *Statistical Analysis; Technical Reports; Testing; *Test Interpretation; *True Scores

ABSTRACT

A new formula is developed for the relative efficiency of two tests measuring the same trait. The formula expresses relative efficiency solely in terms of the standard errors of measurement and, surprisingly, the frequency distributions of true scores. Approximate methods for estimating relative efficiency may make this function routinely available. A numerical illustration compares new and old estimates of relative efficiency for subtests from the Scholastic Aptitude Test. (Author)

RB-73-41

RESEARCH BULLETIN

THE RELATIVE EFFICIENCY OF TWO TESTS AS A FUNCTION OF ABILITY LEVEL

Frederic M. Lord

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
June 1973

THE RELATIVE EFFICIENCY OF TWO TESTS AS A FUNCTION OF ABILITY LEVEL

Abstract

A new formula is developed for the relative efficiency of two tests measuring the same trait. The formula expresses relative efficiency solely in terms of the standard errors of measurement and, surprisingly, the frequency distributions of true scores. Approximate methods for estimating relative efficiency may make this function routinely available. A numerical illustration compares new and old estimates of relative efficiency for subtests from the Scholastic Aptitude Test.

THE RELATIVE EFFICIENCY OF TWO TESTS AS A FUNCTION OF ABILITY LEVEL*

Birnbaum [1968] defines the relative efficiency of two testing procedures as the ratio of their information functions. Their relative efficiency will vary for different levels of the trait measured. Ideally, test manuals should report information functions or relative efficiencies as routinely as they now report reliability coefficients.

The main purpose of the present note is to derive a useful and instructive formula for relative efficiency, appropriate for two unidimensional tests measuring the same trait. It is necessary that the two tests be administered either to the same group or to approximately equivalent groups of examinees. The new formula shows that relative efficiency is closely related to the shapes of the true-score distributions of the two tests.

The first section briefly discusses information functions. The second section derives the new formula. The third section presents a method of practical application and an empirical check.

1. Information Function

A testing procedure produces a score x for each testee, presumed to be related to his standing on the trait θ, hereafter called the "ability," measured by the procedure. The score x may be the number of questions answered correctly, or it may be a complicated function of the examinee's responses. If x were a consistent, preferably an unbiased estimator of θ, and if θ were uniquely defined, the testing and scoring procedure could perhaps be evaluated by its sampling variance. Scores commonly

*Research reported in this paper has been supported by grant GB-32781X from National Science Foundation.

used (because of their convenience) are typically consistent estimators of some awkward function θ , however. Worse yet, this function is seldom the same from one procedure to the next, except for the case, uninteresting for making comparisons, when the two procedures are strictly parallel. This situation usually causes no problems for the mental tester who is interested only in the relative standings of the examinees on θ . For him, within limits, one monotonic function of θ is about as good as another. This situation does prevent us, however, from comparing testing and scoring procedures simply in terms of the sampling variance of the score.

Birnbaum [1968, p. 418] suggests comparing scoring procedures by the widths of their asymptotic confidence intervals for θ . (In this discussion, "asymptotic" indicates that the number, n , of test items is large.) This width is inversely proportional to the square root of

$$(1) \qquad I(\theta,x) \equiv \frac{[\partial \mathcal{E}(x|\theta)/\partial\theta]^2}{\mathrm{Var}(x|\theta)} \quad ,$$

termed the <u>score information function</u>. An alternative, nonasymptotic line of reasoning leading to this same function has been outlined by Lord [1952, eq. 57; 1971, eq. 6.3].

A few remarks about information functions will be listed below:

1. In classical test theory, if x is a linear composite of item scores, lengthening the test k -fold will multiply the mean of x by k . Since $\mathrm{Var}(x|\theta) \equiv \mathrm{Var}[\, x - \mathcal{E}(x|\theta) \mid \theta \,]$ represents the variance of the

errors of measurement, this quantity will be multiplied by k also (not by k^2). Thus, lengthening the test k-fold will multiply the score information function by k. Conversely, a percent increase in a score information function is most easily interpretable as equivalent to the increase achieved by lengthening a conventional test by the same percentage.

2. If x is the maximum likelihood estimator $\hat{\theta}$, then $I\{\theta,x\} \equiv I\{\theta,\hat{\theta}\}$ is asymptotically equal to the Fisher information measure

$$I_F\{\theta,\underset{\sim}{u}\} \equiv \mathcal{E}\left(\frac{\partial \log L\{\underset{\sim}{u}|\theta\}}{\partial\theta}\right)^2 ,$$

where $L\{\underset{\sim}{u}|\theta\}$ is the likelihood function for the vector $\underset{\sim}{u}$ of observed item responses [Birnbaum, 1968, 20.3]. Also, $I\{\theta,\hat{\theta}\}$ is equal to the reciprocal of the asymptotic variance of $\hat{\theta}$.

3. A nonasymptotic line of reasoning given by Rao [1965, pp. 270-1] suggests the use, even for small n, of

$$(2) \qquad I_F\{\theta,x\} \equiv \mathcal{E}\left(\frac{\partial \log L\{x|\theta\}}{\partial\theta}\right)^2$$

as a measure of the information about θ contained in x. This Fisher information measure is necessarily less than or equal to the one given in the preceding paragraph. By virtue of the Cramér-Rao lower bound to the variance, we have under regularity conditions

$$\text{Var } x \leq \frac{[\partial\mathcal{E}(x|\theta)/\partial\theta]^2}{I_F\{\theta,x\}} .$$

Consequently, $I_F\{\theta,\underset{\sim}{u}\} \geq I_F\{\theta,x\} \geq I\{\theta,x\}$. If x is a sufficient statistic for θ, the equality signs hold asymptotically [Kendall & Stuart, 1961, 17.37].

4. A linear transformation of x does not affect $I\{\theta,x\}$, but a nonlinear transformation changes $I\{\theta,x\}$. Asymptotically, the effect of a strictly monotonic nonlinear transformation is negligible under mild conditions.

5. A strictly monotonic nonlinear transformation of x has no effect on the information statistic (2) suggested by Rao, even in small samples, since the likelihood of a sample of observations is not affected by the choice of scoring system. This is a very desirable property, in view of the fact that the choice of a score x rather than some function of x is largely arbitrary. Rao's information measure leads to a very complicated formula, however, when x is the number-right score. For this reason, it will not be utilized here.

6. Let $\theta^* \equiv \theta^*(\theta)$ be a strictly monotonic transformation of the ability scale. It is easily found from the chain rule for differentiation that

$$(3) \qquad I\{\theta^*,x\} = I\{\theta,x\}(\partial\theta^*/\partial\theta)^{-2} \quad ,$$

$$(4) \qquad I_F\{\theta^*,x\} = I_F\{\theta,x\}(\partial\theta^*/\partial\theta)^{-2} \quad .$$

Thus the shape of the information function may be distorted to any continuous single-valued curve by choice of θ^*. In particular, the ability level at which maximum "information" is obtained may be drastically changed by a transformation of the ability scale.

7. It is seen from (3) that the relative efficiency of measuring procedures x and y is <u>not</u> changed by a strictly monotonic transformation of the ability scale. For this reason, the parameter will be omitted from the corresponding symbol:

$$\text{(5)} \qquad \text{R.E.}\{y,x\} \equiv \frac{I\{\theta,y\}}{I\{\theta,x\}} = \frac{I\{\theta^*,y\}}{I\{\theta^*,x\}} \quad .$$

8. Unless we are prepared to defend strongly a particular choice of metric for ability, it will be wise in any practical investigation to present R.E. curves rather than the protean information curves. If desired, an actual measurement procedure can be compared in efficiency to a hypothetical "standard" test composed of statistically equivalent items with specified item parameters, or to a hypothetical standard test characterized by a uniform distribution of item difficulties (Brogden, 1957, p. 305).

2. A New Formula for Relative Efficiency

The relative efficiency of two scores, x and y, is ordinarily computed from their score information functions by (5). As an illustration, consider the case of number-right scores.

For this special case, we have

$$x \equiv \sum_{i=1}^{n} u_i$$

where $u_i = 1$ or 0 represents a right or a wrong answer to item i. Thus

$$\text{(6)} \quad \begin{cases} \mathcal{E}(x|\theta) = \sum_{i=1}^{n} \text{Prob}(u_i = 1|\theta) \\ \qquad\qquad = \sum_{i=1}^{n} P_i(\theta) \quad , \\ \text{Var}\,(x|\theta) = \sum_{i=1}^{n} P_i Q_i \end{cases}$$

where $P_i \equiv P_i(\theta)$ is the characteristic function of item i and

$Q_i \equiv 1 - P_i$. Thus, from (1), the score information function of x for θ is [Birnbaum, 1968, eq. 20.2.2]

$$(7) \qquad I\{\theta,x\} = \frac{\left(\sum_{i=1}^{n} P_i'\right)^2}{\sum_{i=1}^{n} P_i Q_i}$$

where $P_i' \equiv \partial P_i/\partial\theta$. In order to estimate relative efficiencies, it has until now seemed necessary to estimate the item characteristic functions $P_i(\theta)$ for all n_x and n_y items.

Let us now derive a new formula for relative efficiency. We no longer require x to be a number-right score.

By definition, $\xi \equiv \mathcal{E}(x|\theta)$ is the true score corresponding to x . Since $P_i(\theta)$ is ordinarily a strictly increasing function of θ , as will be assumed here, we have from (6) that ξ is also a strictly monotonic transformation of θ . From (3) we then have that the score information function of x for ξ is

$$(8) \qquad I\{\xi,x\} = I\{\theta,x\}(\partial\xi/\partial\theta)^{-2} \quad .$$

Finally, from (1) and (8),

$$(9) \qquad I\{\xi,x\} = 1/\mathrm{Var}(x|\xi) \quad .$$

(The numerator here is 1 because the regression of observed score on true score has unit slope.) If y measures the same trait as x , and $\eta \equiv \mathcal{E}(y|\theta)$ denotes the true score for y , we have similarly

$$I\{\eta,y\} = 1/\mathrm{Var}(y|\eta) \quad .$$

Since ξ and η are both strictly monotonic transformations of θ, it follows that $\eta \equiv \eta(\xi)$ is a strictly monotonic transformation of ξ. Thus we can use (3) to write down the score information function of y for ξ:

$$(10) \qquad I\{\xi,y\} = \frac{(\partial\eta/\partial\xi)^2}{\mathrm{Var}(y|\eta = \eta(\xi))} \quad .$$

The efficiency of y relative to x is now the ratio of (8) and (10):

$$(11) \qquad \mathrm{R.E.}\{y,x\} = \left(\frac{\partial\eta}{\partial\xi}\right)^2 \frac{\mathrm{Var}(x|\xi)}{\mathrm{Var}(y|\eta(\xi))} \quad .$$

Similarly,

$$(12) \qquad \mathrm{R.E.}\{x,y\} = \left(\frac{\partial\xi}{\partial\eta}\right)^2 \frac{\mathrm{Var}(y|\eta)}{\mathrm{Var}(x|\xi)} \quad .$$

The function $\eta(\xi)$ can be defined by the relation

$$(13) \qquad \int_{-\infty}^{\xi_o} p(\xi)d\xi = \int_{-\infty}^{\eta(\xi_o)} q(\eta)d\eta$$

where $p(\xi)$ and $q(\eta)$ are the probability density functions for ξ and η. Equation (13) simply states that for any population, the proportion of cases lying below ξ_o must be the same as the proportion lying

below η_o . This must be true, for ξ and η are simply two different ways of expressing the individual's standing on a single psychological dimension.

Since (13) holds for all ξ_o , we can differentiate both sides to obtain $p(\xi_o) = q(\eta(\xi_o))(\partial\eta_o/\partial\xi_o)$. Dropping the subscript and rearranging gives

$$(14) \qquad \frac{\partial\eta}{\partial\xi} = \frac{p(\xi)}{q(\eta)} \quad .$$

Substituting (14) in (11) gives an interesting expression for the relative efficiency in terms of frequency distributions of true scores:

$$(15) \qquad R.E.\{y,x\} = \frac{Var(x|\xi)}{Var(y|\eta)} \frac{p^2(\xi)}{q^2(\eta)} \quad ,$$

where $\eta \equiv \eta(\xi)$ is the equipercentile equivalent of ξ , as required by (13).

If x is a number-right score, the range of ξ is 0 to n_x , where n_x is the number of items in test x , and similarly for η . It may be desirable to rewrite (15) in terms of $\zeta \equiv \xi/n_x$, $z \equiv x/n_x$, $\omega \equiv \eta/n_y$, and $w = y/n_y$:

$$(16) \qquad R.E.\{y,x\} = \frac{Var(z|\zeta)}{Var(w|\omega)} \frac{g^2(\zeta)}{h^2(\omega)} \quad ,$$

where g and h are the density functions for ζ and ω .

To our surprise, these formulas show that the relative efficiency of two tests can be expressed directly in terms of true-score frequency distributions and standard errors of measurement. The formulas agree with the vague intuitive notion that a test is more discriminating at true-score levels where the scores are spread out, less discriminating at true-score levels where scores pile up.

3. Practical Application

Various convenient ways of estimating the expression on the right of (16) will be found. The crude but simple procedure of substituting sample distributions of observed scores for $p(\xi)$ and $q(\eta)$ will be discussed in another publication. Here we discuss a particular estimation procedure available when x and y are number-right scores. Although this procedure is complicated, it is an order of magnitude simpler than estimating accurately all the item parameters required by (7). In large samples, the new procedure seems to yield results that are much the same for most practical purposes.

The functions $g(\zeta)$ and $h(\omega)$ needed for (16) are estimated from the sample frequency distributions of x and y by methods discussed by Lord [1969], using a revised version, available from the author, of the computer program described by Wingersky, Lees, Lennon, and Lord [1969]. The functions $\mathrm{Var}(z|\zeta)$ needed for (16) are approximated by the formulas [Lord, 1965, eqs. 9, 34]

$$k_x \equiv \frac{1}{2} n_x^2(n_x - 1)s_p^2/[\bar{x}(n_x - \bar{x}) - s_x^2 - n_x s_p^2] \quad , \tag{17}$$

$$\mathrm{Var}(z|\zeta) = (n_x - 2k_x)\zeta(1 - \zeta)/n_x^2 \quad , \tag{18}$$

where $\bar{x}$ and s_x^2 are the sample mean and variance (over people) of the number-right scores, s_p^2 is the sample variance (over items) of the p_i, and p_i is the sample proportion of correct answers to item i. $\mathrm{Var}(w|\omega)$ is obtained similarly.

The relation between η and ξ, symbolized by the notation $\eta \equiv \eta(\xi)$, is calculated numerically by a computer program [Stocking, Lees, Lennon, & Lord, 1969] that solves (13) for η. A revised version of this program, available from the author, also computes relative efficiencies from (16) and plots them as a function of ξ. No item characteristic curve parameters are used anywhere in this method. This new method has been tested out and compared to the old method using 90 verbal items from the Scholastic Aptitude Test. For the old method, item parameters for all 90 items were estimated simultaneously from 2926 answer sheets by the maximum likelihood method described by Lord [1973]. Random responses were supplied for omitted responses (but not for items apparently "not reached" by the examinee). A 45-item "peaked" subtest was selected consisting of those items having estimated difficulty parameters near the average value for the entire test. A "regular" subtest consisted of the 45 even-numbered items. Actually, there was considerable overlap in items between the two subtests. However, the formulas used and the conclusions reached are appropriate for two nonoverlapping 45-item tests having the same item parameters as the actual tests, with all examinees responding to all items.

Estimated score information functions were computed from the estimated item parameters by (7). The dashed curve in Figure 1 shows the ratio of these information functions, estimating the efficiency of the regular

test relative to the peaked test. A logarithmic scale is used for relative eff[illegible]c since an R.E. of .5 is precisely as noteworthy as an R.E. of 2.0.

The solid curve in Figure 1 was obtained by the new method (16), as described in the first paragraphs of this section. The necessary sample distributions for x and y were obtained simply by scoring the 45-item subsets involved. Only the 1805 examinees who finished the test were used for these calculations. The solid curve shows some tendency to oscillate about the first curve, but in general seems to provide a satisfactory and very usable approximation. The oscillations could presumably be avoided by using larger samples or by other means. The computational cost of estimating (16) does not increase with sample size.

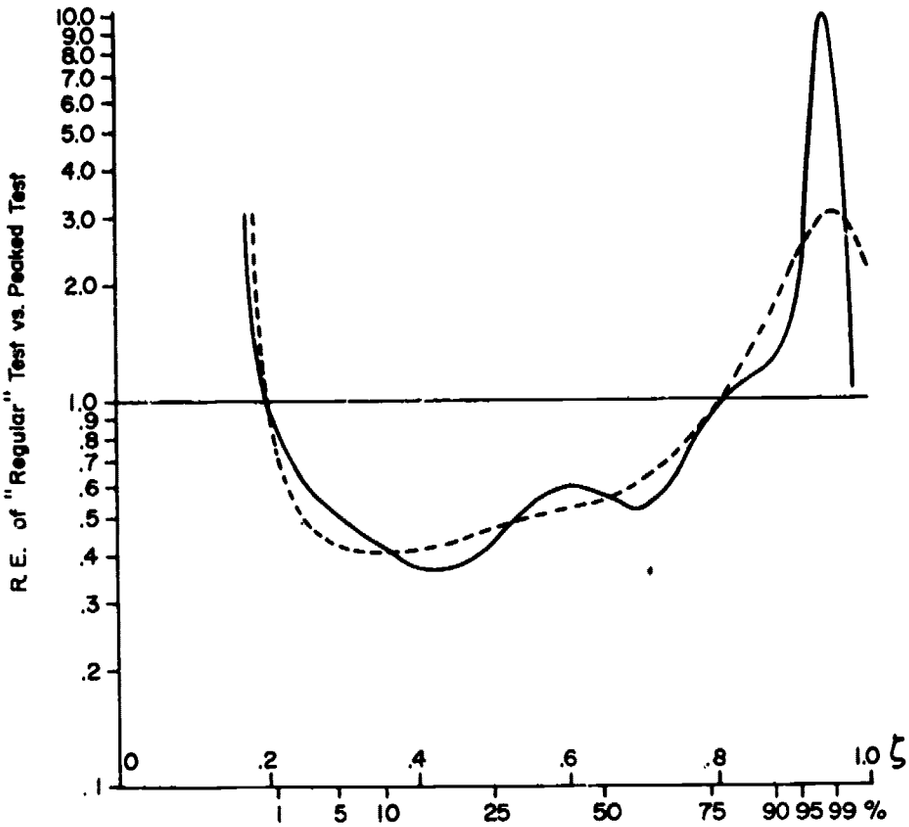

Figure 1. Estimate of relative efficiency from (16) compared with estimate from (7) and (5).

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968. Pp. 453-479.

Brogden, H. E. New problems for old solutions. Psychometrika, 1957, 22, 301-309.

Kendall, M. G. and Stuart, A. The advanced theory of statistics. Vol. 2, New York: Hafner, 1961.

Lord, F. M. A theory of test scores. Psychometric Monograph No. 7. Psychometric Society, 1952.

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

Lord, F. M. Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). Psychometrika, 1969, 34, 259-299.

Lord, F. M. Tailored testing, an application of stochastic approximation. Journal of the American Statistical Association, 1971, 66, 707-711.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Research Bulletin 73-37. Princeton, N.J.: Educational Testing Service, 1973.

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

Stocking, M., Lees, D. M., Lennon, V. and Lord, F. M. A program for estimating a bivariate distribution of test scores from the marginals. Research Memorandum 69-2. Princeton, N.J.: Educational Testing Service, 1969.

Wingersky, M. S., Lees, D. M., Lennon, V. and Lord, F. M. A computer program for estimating true-score distributions and graduating observed-score distributions. Educational and Psychological Measurement, 1969, 29, 689-692.