THE RELIABILITY OF TEST DISCRIMINATIONS[1]

JOHN E. MILHOLLAND
University of Michigan

I. *Introduction*

SEVERAL recently developed methods of administering multiple choice test items have possibilities for making more discriminations among individuals by means of a given number of items. The *SRA Self-Scorer* (1), devised by Angell and Troyer, requires the subject to punch holes in an answer sheet. If he punches a hole corresponding to the answer to the item, he uncovers a red dot; otherwise he uncovers a white dot, and he goes on punching until he does uncover a red dot. Thus, for an item with k alternatives, possible scores, (i.e., number of holes punched) range from 1 to k instead of from 0 to 1 as is the case with items conventionally administered. Angell and Troyer refer to a forerunner of this procedure, a machine constructed by Pressey (8), and described in 1927.

C. H. Coombs (2) has proposed instructing examinees to indicate as many of the k-1 item distracters as they are sure are distracters. The scoring formula assigns a score of 1 to every distracter identified, and 1-k to an answer falsely identified as a distracter. Each item, then, has a score range from 1-k to k-1. In a four-choice item, for example, the score range would be from -3 to $+3$.

Dressel and Schmid (4) have experimented with several methods of presentation, one of which is formally the same as that of Coombs. Another of their procedures required that the examinee register his confidence in the correctness of his answer.

[1] Support for this research was provided by a Summer Faculty Research Fellowship granted by the Horace H. Rackham School of Graduate Studies, University of Michigan. Indebtedness is also acknowledged to C. H. Coombs for pertinent suggestions and helpful discussions.

Items were then scored in accordance with the following schedule:

Degree of Confidence	Item Score	
	If Right	If Wrong
Positive	+4	−4
Fairly certain	+3	−3
Rational guess	+2	−2
No defensible basis for choice	+1	−1

Thus if a student chose an incorrect alternative as the answer and was "fairly certain" it was correct, his score on the item would be −3.

A method similar to this has been used by Dr. Philip Nogee[2] at Boston University. After the student made his choice for the answer to an item he selected one of five fractions (in the case of a 4-choice item) for a scoring formula. The fractions were: 0/0, 1/.33, 2/.67, 3/1, and 4/1.33, with the numerator the number of points the student received if his answer was right, the denominator the number he was docked if he was wrong.

Another variant has been reported by Leichty (7). He required students to select an answer and then judge each alternative (including the one selected as the answer) as belonging to one of the following categories:

1. Wholly true and relevant to the problem of the item
2. Partially true and relevant
3. Wholly false, but relevant
4. True, but irrelevant
5. False and irrelevant

Although in the cases of several of the above methods the objectives of the originators did not explicitly include increasing the ability of the test to make discriminations among individuals, it can be seen that they all have this potential. The question may arise, then, as to whether the additional discriminations are reliable or are merely artificial distinctions with a large error component. The development of procedures for assessing the reliabilities of test discriminations is the concern of this paper.

[2] Personal communication.

The term "discrimination" appears frequently in the literature of test theory and development, but it is almost always used in the context of item selection. Ferguson, however, has devised an index of test discrimination (5) which is the ratio of the number of between-persons discriminations actually made by a test to the maximum possible for the same sample size and number of score categories. This same index was independently derived by Thurlow (9), who also considered the matter of reliability. He suggested that, given one or more retests, a difference between two persons be considered reliable if it is, on the average, larger than the differences between retest scores belonging to each individual of the pair.

The "sensitivity" ratio of Jackson (6) could also be used as a measure of the reliability of test discriminations. This index is the ratio of the true-score standard deviation to the error standard deviation. Standards of magnitude for it could be set so as to place limits on the relative size of errors of measurement.

The remaining sections of this paper will present some additional procedures for attacking the problem of reliable discriminations.

II. *The Number of Statistically Reliable Discriminations*

The question of discrimination reliability may be approached through the consideration of the proportion of discriminations made by the test which are based on score differences which exceed some criterion magnitude—the five per cent confidence level, say, of the distribution of errors of measurement. When the form of the distribution of differences is known, or assumed, this requirement can be expressed in terms of Jackson's sensitivity index. Some more general implications of this approach will be indicated later.

For a score distribution with variance σ_t^2, the distribution of the differences between all pairs of scores (including the zero differences between individuals and themselves) will have a mean 0 and variance $2\sigma_t^2$. The standard error of the difference between two scores is given by $\sigma_t\sqrt{2(1 - r_{11})}$, where r_{11} is the reliability coefficient of the test.

If $f(x)$ is the probability density function of the differences between scores, and k standard errors defines the desired significance level, the proportion of significant differences is

$$p = 2\int_{k\sigma_t\sqrt{2(1 - r_{11})}}^{\infty} f(x)dx \,.$$

Changing to units of the standard deviation of the distribution of differences by letting $\sigma_t\sqrt{2} = 1$ and $g(y)$ the density function gives

$$p = 2\int_{k\sqrt{1 - r_{11}}}^{\infty} g(y)dy \,.$$

For a particular k, this proportion is dependent entirely upon the reliability coefficient of the test and the characteristics of the distribution of differences. The implication of this result is that, for a given form of the difference distribution, the test reliability must be increased in order to obtain a larger number of significant discriminations. This is a consequence of counting all comparisons between individuals, including zero differences, as discriminations. For a given number of people, N, there are always N^2 discriminations, only N of which are actually considered to be zero. In the real case, of course, it is to be expected that a test with more score categories will have fewer zero differences, but, again, this is not likely to affect the extremes of the difference distribution, where the significant differences lie.

This dependence of statistically reliable discriminating ability upon the reliability coefficient is also illustrated by expressing Jackson's sensitivity measure in the form.

$$\gamma = \sqrt{\frac{r_{11}}{1 - r_{11}}}\,.$$

III. *Comparative Reliability of Maximal Discrimination*

Ferguson (5) has shown that the number of between-persons discriminations made by a test is maximum when the distribution of scores is rectangular. The question of whether additional discriminations engendered by an expanded score range would be valuable may be considered in terms of the reliability coefficient necessary so that a difference of one score point is as reliable in the expanded scores as in the original scores. This condition may be expressed as the requirement that the standard errors of measurement of both sets of scores be the same.

The variance of a rectangular distribution for n items scored $1 - 0$, so that there are $n + 1$ score categories, is given by $\frac{n(n+2)}{12}$. If, now, the items are scored 0 through a, so that there are $an + 1$ score categories, the variance is $\frac{an(an+2)}{12}$. Equating error variances, with r_{11} representing the reliability coefficient when the items are scored $1 - 0$, r_{aa} the reliability coefficient of the expanded test, gives

$$\frac{n(n+2)}{12}(1 - r_{11}) = \frac{an(an+2)}{12}(1 - r_{aa}).$$

Solving for r_{aa}:

$$r_{aa} = 1 - \frac{(n+2)(1 - r_{11})}{a(an+2)} = 1 - \frac{(n+2)(1 - r_{11})}{a^2\left(n + \frac{2}{a}\right)}.$$

If n, the number of items, is fairly large, i.e., 25 or so, the fraction $\frac{n+2}{n+\frac{2}{a}}$ will be fairly close to 1, so that

$$r_{aa} \doteq 1 - \frac{1}{a^2}(1 - r_{11}) \qquad (1)$$

could be used for an estimate of r_{aa}.

If increasing the range of scores from $n + 1$ to $an + 1$ has the effect of multiplying the standard deviation of the score distribution, whatever its form, by a factor a, equating the two standard errors of measurement would give

$$\sigma_1^2(1 - r_{11}) = a^2\sigma_1^2(1 - r_{aa}),$$

where σ_1 is the standard deviation of the original test, before expansion. This equation solved for r_{aa} is

$$r_{aa} = 1 - \frac{1}{a^2}(1 - r_{11})$$

and this is the same as the approximation formula (1).

TABLE 1

Reliabilities Required for a Constant Standard Error of Measurement Compared with Those Predicted by the Spearman-Brown Formula, for an Expansion Factor of 3

Original Test Reliability Coefficient	Approximate Coefficient Required for Constant σ_{meas}	Coefficient Predicted by Spearman-Brown Formula
.50	.94	.75
.60	.96	.82
.70	.97	.88
.80	.98	.92
.90	.99	.96

It may be seen that the requirement of equal standard errors of measurement would be a severe one to impose. This is shown in Table 1, where, taking the expansion factor as 3, some required reliabilities are compared with those predicted by the Spearman-Brown formula for a test three times as long. Even if the new response methods described in the Introduction result in an improvement in reliability corresponding to a lengthening of the test by the same factor by which they expand the score range (and there is some evidence (3) that the effect is less than this), much of the added discriminating power must be considered valueless.

IV. *The Expected Proportion of Discriminations Which Represent True Differences*

The discussions of Section II and III have illustrated the nature of the dependence of discrimination reliability on test reliability. The present section is devoted to a consideration of what proportion of the differences of any given size may be expected to reflect true differences in the same direction.

If it is assumed that errors of measurement in the difference distribution are random and normally distributed with mean 0 and variance $2\sigma_t^2(1 - r_{11})$ for every score difference, it is possible to compute the expected proportion of non-error differences for any particular difference in scores. The procedure involves merely determining the portion of the positive tail of the error distribution lying beyond the abscissa value corresponding to the difference in question. The formula for the error abscissa corresponding to a difference of k standard deviations of the score distribution is

$$\frac{x}{\sigma_e} = \frac{k}{\sqrt{2(1 - r_{11})}}.$$

TABLE 2

Expected Percentages of True Differences Greater than Zero for Various Reliability Coefficients and Magnitudes of Obtained Differences

Reliability Coefficient	Obtained Test z-Score Difference = Raw Difference/σ Test																					
	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2
.99	76	92	98	99.8																		
.98	69	84	93	98	99																	
.97	66	79	89	95	98	99																
.96	64	76	86	92	96	98	99															
.95	63	74	83	90	94	97	99															
.94	61	72	81	87	93	96	98	99														
.93	61	70	79	86	91	95	97	98	99													
.92	60	69	77	84	89	93	96	98	99													
.91	59	68	76	83	88	92	95	97	98	99												
.90	59	67	75	81	87	91	94	96	98	99												
.89	58	67	74	80	86	90	93	96	97	98	99											
.88	58	66	73	79	85	89	92	95	97	98	99											
.87	58	65	72	78	84	88	91	94	96	98	98	99										
.86	58	65	72	78	83	87	91	93	96	97	98	99										
.85	57	64	71	77	82	86	90	93	95	97	98	99										
.83	57	63	70	75	81	85	88	91	94	96	97	98	99									
.81	56	63	69	74	79	83	87	90	93	95	96	97	98	99								
.79	56	62	68	73	78	82	86	89	92	94	96	97	98	98	99							
.77	56	61	67	72	77	81	85	88	91	93	95	96	97	98	99							
.75	56	61	66	72	76	80	84	87	90	92	94	96	97	98	98	99						
.70	55	60	65	70	74	78	82	85	88	90	92	94	95	96	97	98	99					
.65	55	59	64	68	72	75	79	82	85	87	90	92	93	95	96	97	98	98	99			
.60	54	59	63	67	71	75	78	81	84	87	89	91	93	94	95	96	97	98	98	99		
.55	54	58	63	66	70	74	77	80	83	85	88	90	91	93	94	95	96	97	98	98	99	
.50	54	58	62	66	69	73	76	79	82	84	86	88	90	92	93	95	96	96	97	98	98	99

The area under a unit normal curve beyond this value gives the proportion of errors large enough to reverse the direction of the obtained difference. The complement of this proportion, then, is the expected proportion of true differences which are in the same direction as the obtained difference. Proportions for various reliability coefficients and magnitudes of difference are shown in Table 2.

V. *Summary*

The answer to the original question which prompted this investigation, viz: whether additional discriminations made by various modified response methods for multiple choice items are reliable, seems to be negative, unless the modified methods improve the reliability coefficient of the test. This result follows if reliability of discrimination is defined in terms of the proportion of differences between scores which attain a given level of statistical significance.

Supplementary developments have included:

1. The derivation of a formula for computing the reliability coefficient necessary for a given score difference on a test with expanded score range to be as reliable as that same difference on a test with restricted score range, assuming both tests to be at maximum discriminating power.

2. The construction of a table giving, for various magnitudes of differences between scores and for various reliability coefficients, the expected proportions of true differences in the obtained direction.

REFERENCES

1. Angell, G. W., and Troyer, M. E. *SRA Self-Scorer.* Chicago: Science Research Associates, 1949.
2. Coombs, C. H. "On the Use of Objective Examinations." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XIII (1953), 304–310.
3. Coombs, C. H., Milholland, J. E., and Womer, F. B. "The Assessment of Partial Knowledge in Objective Testing." *PRB Technical Research Note 33*, Department of the Army, 1955.
4. Dressel, P. L., and Schmid, J. "Some Modifications of the Multiple-Choice Item." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XIII (1953), 574–595.

5. Ferguson, G. A. "On the Theory of Test Discrimination." *Psychometrika*, XIV (1949), 61–68.
6. Jackson, R. W. B. "Reliability of Mental Tests." *British Journal of Psychology*, XXIX (1939), 267–287.
7. Leichty, V. E. "Student Thinking on Short Answer Examinations." *Journal of Educational Research*, XLIII (1949), 41–48.
8. Pressey, S. L. "A Machine for the Automatic Teaching of Drill Material." *School and Society*, XXV (1927), 549–552.
9. Thurlow, W. R. "Direct Measures of Discrimination among Individuals Performed by Psychological Tests." *Journal of Psychology*, XXIX (1950), 281–314.