DOCUMENT RESUME

ED 201 665 TM 810 258

AUTHOR Frisbie, David A.
TITLE A Method for Comparing Test Difficulties.
PUB DATE Apr 81
NOTE 12p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Los Angeles, CA, April 11-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Comparative Testing; *Difficulty Level; *Evaluation Methods; *Raw Scores; *Scaling; Test Items; *Transformations (Mathematics)
IDENTIFIERS *Relative Difficulty Ratio

ABSTRACT

The relative difficulty ratio (RDR) is used as a method of representing test difficulty. The RDR is the ratio of a test mean to the ideal mean, the point midway between the perfect score and the mean chance score for the test. The RDR tranformation is a linear scale conversion method but not a linear equating method in the classical sense. The transformation is used with non-parallel tests, tests which may differ in both item format and length. The goal of the transformation is not to "equate" for differences in test difficulty. The purpose of the RDR transformation is to convert the raw score scale from a given test so that the new scale has the same mean chance point and the same range or length as the raw score scale from a second test. The goal is to eliminate differential guessing chance factors and unequal test lengths as competing explanations for why two test means are different (or the same). A method for transforming one of the raw score scales so that an inferential test of mean differences can be accomplished though the original raw score scales vary in mean chance score and/or length is described. (RL)

A METHOD FOR COMPARING TEST DIFFICULTIES

David A. Frisbie

Measurement and Research Division
Office of Instructional Resources
University of Illinois at Urbana-Champaign

Presented at the Annual Meeting of the
National Council on Measurement and Education

Los Angeles
April 1981

Introduction

How can the difficulties (means) of two content-parallel tests be compared if such tests differ in mean chance scores or test lengths? For example, the means of a 100-item four-choice multiple choice test and a content-parallel 150-item true false test cannot be compared meaningfully. Their chance scores are 25 and 75, respectively, and their lengths are in a two to three ratio. When administered to the same or randomly equivalent groups, test reliabilities and validity can be studied directly, but test difficulty cannot be addressed using the mean raw scores.

When faced with this problem, some researchers have made inappropriate comparisons which have yielded inaccurate conclusions while others have simply identified the lack of comparability without drawing conclusions. Studies by Ebel (1978), Frisbie (1973, 1974), and Mendelson, et al. (Note 1) are examples of research aimed at discovering differences in properties of content-parallel tests in which item format varied. A study by Hughes and Trimble (1965) illustrates how some researchers vary the nature of distractors in multiple choice items which also vary in the number of choices per item. Item format has been used as an independent variable in studies where another independent variable may be confounded with item format. Benson and Crocker (1979) and Huck (1978) both represent examples.

In view of the obstacles to comparison presented by chance score differences and variability in test length, Frisbie (in press) proposed the use of a relative difficulty ratio (RDR) as a method of representing test difficulty. The RDR is simply the ratio of a test mean to the ideal mean, the point midway between the perfect score and the mean chance score for the test. The multiple choice and true-false tests referred to above would have ideal means of 62.5 and 112.5.

These means are ideal in the sense that, for norm-referenced purposes, the overall discrimination capabilities of the tests are maximized. A convenient scale transformation can be applied to the basic definitional formula of RDR to arrive at the computational formula which appears as equation one. The value of RDR using equation one, where $\bar{X} = 70.5$, $K = 100$, $\bar{X}_I = 62.5$, K is the number of items,

$$RDR = (\bar{X} - \bar{X}_I)\ (K - \bar{X}_I) \qquad (1)$$

and $\bar{X}_I$ is the ideal mean which is .213. Positive values of RDR are associated with tests which are too easy and negative values are associated with tests which are too difficult in the norm-referenced sense. The RDR for two tests which are comparable in content could be compared to determine the relative difficulty of either.

A significant limitation of using the RDR statistic for judging test difficulty is that sampling fluctuation is ignored. The RDR is strictly descriptive; probability statements about observed differences being statistically significant cannot be made because the theoretical random sampling distribution of RDR has not been identified. Another limitation of using RDR to make inferences is that the RDR scale is an uncommon one; RDR differences are less easily interpreted than raw score differences.

The purpose of this paper is to describe a method for transforming one of the raw score scales so that an inferential test of mean differences can be accomplished, though the original raw score scales vary in mean chance score and/or length. The transformation using RDR overcomes the limitations cited above of using RDR alone to make inferences. In addition, though the transformation is linear, it does not accomplish linear equating as such equating is classically defined.

Theoretical Development

For purposes of illustration, we assume that two tests, A and B, have been developed so as to be content parallel in the general or specific sense (i.e., items were written from sampling the same population of instructional objectives or items were written in one format and then converted to another format). Furthermore, A and B are choice-type tests with different mean chance scores (i.e., the number of choice options varies) and unequal test lengths. The problem is to statistically test the difference between the means of A and B, derived from administering A and B to the same group or randomly equivalent groups. To do so, the scores on Test A must be converted to the Test B score scale, or vice versa. Then we can choose the appropriate statistical technique (e.g., t-test, ANOVA) in normal fashion for testing statistical significance.

The Transformation

The scale transformation begins with equation 2, the RDR of the test scale on which the transformation is to be performed, Test A. The RDR for Test B can be

$$RDR_A = (\bar{X}_A - \bar{X}_{IA})/(K_A - \bar{X}_{IA}) \qquad (2)$$

expressed in identical fashion by using the B subscript instead of A. RDR_B can be rewritten to solve for $\bar{X}_B$ as shown in equation 3.

$$\bar{X}_B = RDR_B\ (K_B - \bar{X}_{IB}) + \bar{X}_{IB} \qquad (3)$$

If test B were to yield an RDR equal to RDR_A, what would the mean of Test B equal? The mean of Test A can be transformed to the Test B score scale by substituting RDR_A in equation 3 for RDR_B. Equation four represents the transformed mean, $\bar{X}'_B$, where the prime denotes a transformed value.

$$\bar{X}'_B = RDR_A\ (K_B - \bar{X}_{IB}) + \bar{X}_{IB} \qquad (4)$$

To arrive at a computational formula for $\bar{X}'_B$, equation two is substitut equation four to yield

$$\bar{X}'_B = \frac{(\bar{X}_A - \bar{X}_{IA})}{(K_A - \bar{X}_{IA})}(K_B - \bar{X}_{IB}) + \bar{X}_{IB}$$

Through multiplication and algebraic simplification equation five can be wri as shown in equation six.

$$\bar{X}'_B = \bar{X}_A \left[\frac{(K_B - \bar{X}_{IB})}{(K_A - \bar{X}_{IA})}\right] + \frac{(K_A\bar{X}_{IB} - K_B\bar{X}_{IA})}{(K_A - \bar{X}_{IA})}$$

Equation six is written to show that the transformation is linear; the mean of Test A is multiplied by a constant and an additional constant i added to that product. All scores on Test A can be converted to the Test B score scale via equation six by replacing $\bar{X}'_B$ with X'_B and $\bar{X}_A$ with X_A. Equat seven shows that the standard deviation for Test A can be converted to the B scale by dropping the second term of equation six.

$$S'_B = S_A \left[\frac{(K_B - \bar{X}_{IB})}{(K_A - \bar{X}_{IA})}\right]$$

The mean score difference, $\bar{X}_B - \bar{X}'_B$, can be tested for statistical sign ance using whatever procedure would be appropriate had A and B not had differen mean chance scores or test lengths. The appropriate standard error can be derived from S_B and S'_B or from the Test B raw scores and the transformed Test A raw scores.

Assumptions

The assumptions made in transforming $\bar{X}_A$ to $\bar{X}'_B$ include those associated with RDR (Frisbie, in press): (1) a wrong response to an item is a random guess from options in that item, (2) dichotomous scoring is used, and (3) there is only one correct answer per item. In addition, if Test B is longer than Test A, the transformation of X_A to X'_B requires that the assumptions associated with the

familiar Spearman-Brown Prophecy Formula be applied to the test. That is, the theoretically-lengthened test adds items which are equivalent in content and difficulty to the original items and the added length would not contribute to examinee fatigue or change in examinee psychological set.

Though these assumptions are commonly made in psychometric research, the first, regarding random guessing, is troublesome to many measurement specialists. Many would say that the assumption is unreasonable because, in practice, it is always violated. Yet in making item format comparisons using RDR, this assumption is applied systematically to both tests being compared. There is little reason to expect that the effect of violating the random guessing assumption would be much more influential on one item format than another.

Further assumptions may need to be made, depending on the variability of scores and number of items on the raw score scales to be compared. For example, negative numbers may appear on the transformed scale if highly variable scores on a long scale are transformed to a relatively short scale. An interval measurement scale must be assumed in such cases. Negative score values can be avoided by applying the transformation to the set of scores having the smallest amount of variability. Non-integer values of transformed scores should be retained so that precision is not lost if statistical tests are performed.

Discussion

The RDR transformation is a linear scale conversion method but not a linear equating method in the classical sense. The transformation is used with non-parallel tests, tests which may differ in both item format and length. The goal of the transformation is not to "equate" for differences in test difficulty.

The conversion process is not intended to yield equivalent scores as one might accomplish by employing linear or equipercentile equating (Anghoff, 1971). The purpose of the RDR transformation is to convert the raw score scale from a given test so that the new scale has the same mean chance point and the same range or length as the raw score scale from a second test. The goal is to eliminate different guessing chance factors and unequal test lengths as competing explanations for why two test means are different (or the same). Though some researchers have used a correction for guessing to address the chance score problem, the test length problem still remained unresolved. The correction for guessing solution is generally a less attractive solution because, if examinees are informed about the correction, risk-taking behaviors and differential guessing or omitting strategies tend to introduce sources of score invalidity. The use of the RDR transformation requires no special scoring and no special directions to examinees and it accounts for test length differences in the measures to be compared.

The data in Table 1 are artificial scores used to illustrate an application of the RDR transformation. Assume that Test A has 12 four-choice multiple choice items, Test B has 18 true-false items, and that these tests were administered

[Insert Table 1 About Here]

to chance halves of a group of 20 examinees. Ignoring the RDR transformation, a t-test of the difference, $\bar{X}_A - \bar{X}_B$, is significant at $\alpha < .005$ with nine degrees of freedom. This result leads to the erroneous conclusion that Test A is more difficult than Test B.

On a purely descriptive basis, the RDRs shown in Table One indicate that Test A is slightly easier than Test B. The equation at the bottom of the table is used to calculate the mean for Test A as it would appear on the Test B scale. When the difference, $\bar{X}'_B - \bar{X}_B$, is tested for statistical significance, the result ($t = 0.261$) is not significant at a reasonably acceptable level.

The conclusion that Tests A and B do not differ in difficulty for the population under consideration is warranted because the mean chance score difference and the test length difference have been adjusted by the RDR transformation.

The adjustments made by the RDR transformation are essential when investigating the effect of item format on test difficulty. If such an effect exists, then its impact on validity must be determined. Most measurement practitioners assume that objective item formats are interchangeable; i.e., it makes no difference if multiple choice, true-false, or matching items are used to measure achievement. If these formats do have a differential effect on validity, then the circumstances under which each is optimumly valid need to be investigated. Such studies could begin by applying the RDR transformation to item difficulties. The formula for item RDRs (Frisbie, in press) can be used to derive item level calculations analagous to equations three through five presented above. If item RDRs were to yield important differences, case study methods might be useful for investigating the test taking process variables which contribute to these differences.

Table 1

Illustrations of the RDR Transformation

Statistics	Test Scores		
	X_A	X'_B	X_B
	10	16.4	18
	10	16.4	18
	10	16.4	18
	9	15.7	17
	9	15.7	16
	8	14.9	16
	8	14.9	15
	7	14.1	13
	7	14.1	12
	7	14.1	7
K	12	18	18
$\bar{X}$	8.5	15.3	15
S_x	1.26	0.976	3.50
RDR	.222	.222	.143
$\bar{X}'_B$	(.222) (3.5) + 14.5		

Reference Notes

1. Mendelson, M. A., et al. The effect of format on the difficulty of multiple-completion test items. Paper presented at the meeting of the National Council on Measurement in Education, Boston, April 1980.

References

Anghoff, W. H. Scales, norms, and equivalent scores. In Thorndike, R. L. (Ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Benson, J. & Crocker, L. The effects of item format and reading ability on objective test performance: A question of validity. Educational and Psychological Measurement, 1979, 39, 381-387.

Ebel, R. L. The ineffectiveness of multiple true-false test items. Educational and Psychological Measurement, 1978, 38, 37-44.

Frisbie, D. A. Multiple choice vs. true-false: A comparison of reliabilities and concurrent validities. Journal of Educational Measurement, 1973, 10, 297-304.

Frisbie, D. A. The effect of item format on reliability and validity: A study of multiple choice and true-false achievement tests. Educational and Psychological Measurement, 1974, 34, 885-892.

Frisbie, D. A. The relative difficulty ratio--a test and item index. Educational and Psychological Measurement, in press.

Huck, S. W. Test performance under the condition of known item difficulty. Journal of Educational Measurement, 1978, 15, 53-58.

Hughes, H. H. & Trimble, W. E. The use of complex alternatives in multiple choice items. Educational and Psychological Measurement, 1965, 25, 117-125.