DOCUMENT RESUME

ED 206 713	TM 810 616
AUTHOR	Carsrud, Karen Banks; Ligon, Glynn
TITLE	Equating Studies: A Manual of Issues, Options and Decisions for Public School Evaluators.
INSTITUTION	Austin Independent School District, Tex.
REPORT NO	AISD-80-50
PUB DATE	Apr 81
NOTE	11p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Cutting Scores; *Equated Scores; *Methods; Predictive Measurement; Statistical Analysis; Test Construction

ABSTRACT

Based on the experiences of four equating studies conducted by the Austin (Texas) Independent School District, a practical "cookbook" approach to test equating is presented. Three tvpes of equating procedures are discussed: choosing a cutoff score on a new instrument, predicting Y from X, and symmetric equating of X and Y. (BW)

Equating Studies: A Manual of Issues, Options, and Decisions for Public School Evaluators

Karen Banks Carsrud and Glynn Ligon
Austin Independent School District

This paper was presented at the annual convention of the American Educational Research Association, Los Angeles, April, 1981.

Publication: #80.50

Equating Studies: A Manual of Issues, Options, and Decisions for Public School Evaluators

Karen Banks Carsrud and Glynn Ligon
Austin Independent School District

The theoretical works and computational formulas previously published in the area of test equating proved less helpful than expected on several occasions when the Office of Research and Evaluation of the Austin Independent School District found a need to adopt new tests. In each case, it was important that the eligibility requirements for students needing special programs remain constant, and/or that the data collected with the new instrument be comparable to longitudinal data collected with the previous instrument. Although various relevant technical references were found, practical topics such as sampling and test administration were not discussed in a step-by-step manner in these references. A practical manual or "cookbook" approach to conducting test-equating studies was apparently needed.

Multiple or parallel test forms invariably differ in terms of difficulty level and score range (Jaeger, 1980; Angoff, 1971). These differences in range and difficulty level are even more apparent between different tests which purport to measure the same dimension, but are not parallel in development. Thus, the public school evaluator and other persons who are involved in measurement using tests will eventually encounter the need to equate scores on multiple tests or multiple forms.

Angoff (1971) states that a commonly accepted definition of equivalent scores is:

> "Two scores, one on Form X and the other on Form Y (where X and Y measure the same function with the same reliability), may be considered equivalent if their corresponding percentile ranks in any given group are equal (page 563)."

Using this definition, it can be argued that tests which are not truly parallel or unidimensional cannot be equated. However, it may often be necessary to attempt this process, even when assumptions of equivalency are not met. Score conversions may be crucial to the usefulness of any test once another test is developed to measure the same trait.

The issues and suggestions that will be the focus of discussion in this paper arose from the experiences of the Office of Research and Evaluation in the Austin Independent School District in conducting four equating-type studies. Briefly, the four studies were concerned with:

1) equating of related subtests on Levels 7 - 14 of the 1978 Iowa Tests of Basic Skills and Levels 1 - 4 of the 1970 California Achievement Test (Ligon and Matter, 1980);

2) choosing a cutoff score on the Comprehensive English Language Test that is equivalent to an existing cutoff on the Bilingual Syntax Measure (Ligon and Matusek, 1978; Matusek and Ligon, 1980);

3) determining the cutoff scores on forms A and B of the Sequential Tests of Educational Progress that are equivalent to the state competency standards on the 1980 Texas Assessment of Basic Skills (Baenen and Curtis, 1980); and

4) determining the cutoff score on the Texas Assessment of Basic Skills which would be equivalent to the 1980 Austin Independent School District graduation requirements based on the Sequential Tests of Educational Progress (Baenen and Curtis, 1980).

Three types of equating procedures will be discussed in this paper. The last three studies mentioned above are examples of a special case of equating: choosing an "equivalent" cutoff on a new instrument. The other two equating procedures discussed involve equating scores along the full range of scores on X and Y, as represented by the first study above.

$$\underline{X_c = Y_c}$$

Choosing a Cutoff Score on a New Instrument

Introduction. Many tests are administered primarily in order to determine whether a student has reached a certain proficiency level. For example, minimum competency tests, language proficiency tests of limited-English-proficiency (LEP) students, and certain tests of basic skills have a cutoff or minimum scores that a student must reach in order to graduate, exit from LEP status, or be promoted to the next grade.

Inevitably, such tests are either revised or become outdated and are replaced with a new test or a new version of the old test. The problem of choosing a new cutoff score on the new test then arises.

Considerations. Most tests come with norms that include percentile, stanines, or grade equivalents. However, one cannot be sure that a raw score corresponding to the 50th percentile on a test normed in 1970 will be truly equivalent to a raw score corresponding to the 50th percentile on a test normed in 1980. Often, the cutoff score on the new instrument (Y) is intended to be equivalent to the cutoff score on the old instrument (X), rather than correspond to some absolute normative criterion.

In addition to having normative samples drawn at two different points in time, the samples may also contain a different ethnic balance, or in some cases, norms may not be provided at all. In short, norms provided with the two tests may not prove useful in establishing a cutoff on the new test.

There are basically two types of classification errors to consider in choosing a new cutoff. The first concerns students who would not have met the criterion using the old test (X) and cutoff, but do meet the criterion on the new test (Y)--false "passes." The second type of error is concerned with students who would have met the criterion on the old test and cutoff (X), but do not meet the criterion on the new test and cutoff (Y)--false "failures." Choice of the new cutoff score on Y should consider the implications of each of these two types of error.

In some cases, the false "passes" (or students who reach the criterion on Y but would not have reached the criterion on X) would no longer be eligible for some special compensatory service (such as a competency tutorial or a bilingual program). In such a case, it would be undesirable to set a cutoff that resulted in too many false "passes" and removed students from programs that were still needed.

In other cases, a false "failure" might prevent a student from qualifying for an accelerated program or graduating on time. Determining the relative importance of each type of error is a major step in choosing the new cutoff score.

The choice of a cutoff score on a new test or test form that is equivalent to a pre-existing cutoff score on another test or test form is a special case of deriving equivalent scores. For this special case, this paper will suggest an equating technique that does not equate scores along the full range of scores on the two instruments (Guilford, 1965; Matusek and Ligon, 1980).

Suggested Steps

1) Determine the relative importance of the two types of classification errors ("false passes" and "false failures"), and the maximum acceptable rates for each type of error.

2) Sampling and administration: Remember that the study is not designed to equate X and Y along the entire range of scores. Therefore, the most efficient use of subjects would be to choose a sample for testing with Y for which scores on X ranged about the cutoff on X. Three problems occur with this approach. First, it is generally preferable to counterbalance the order of administration when equating tests in order to minimize systematic effects due to practice and fatigue. Second, scores on X are not always available in advance, and thus, it would not be possible to choose subjects whose scores ranged about a cutoff on X. (However, if recent scores on X are already available, retesting on X and counterbalancing administration of X with Y may be inefficient and also result in inflated scores on X due to repeated testing.)

A third problem arises from the techniques used for data analysis. The procedure suggested below and by Guilford (1965) assumes that scores on Y are normally distributed, and that the proportion of passes and failures on X in the sample would be the same as for the population. If the second assumption concerning the sample and population proportions of passes and failures is met, truncating the tails of the distribution on Y is probably not a serious violation of the normality assumption, and may be a more efficient use of subjects. However, the evaluator must still consider whether efficiency is a more important consideration than counterbalancing the order of administration of the instruments.

If scores on X are known and the evaluator is concerned about the most efficient possible use of subjects, the following procedure may be helpful. First, determine the largest sample that would be feasible for the study. Then, determine the actual number of persons in the sample who should fall above and below the cutoff on X. For example, if 20 percent of the population fall above the cutoff, 40 persons in a sample of 200 should be above the cutoff. Finally, using the example above, the 160 persons in the population who score immediately below the cutoff and the 40 persons who score immediately above the cutoff would be administered Y.

Analyses:

3) Choose a preliminary cutoff to minimize overall errors of classification, using the formula suggested by Guilford (1965; page 385):

$$y_h = M_y + \left(\frac{yz}{pq}\right)\left(\frac{\sigma_y^2}{M_p - M_q}\right)$$

y_h = the critical value on y

M_y = Mean of all y values

p = Proportion of cases passing on X

q = 1 - p

M_p = Mean of y values for proportion passing on X

M_q = Mean of y

σ_y^2 = Variance in the total distribution of y

y = Ordinate in the unit normal distribution at the point of division of the area under the curve with p proportion above it

z = standard measure of the point at which the division occurs

4) Determine the percentage of each type of misclassification resulting from use of the preliminary cutoff score:

		X	
		PASS	FAIL
Y	PASS		?
CUTOFF:____	FAIL	?	

5) Based on the type of classification error that is least desirable (false failures versus false passes), determine the cutoff score on Y that would eliminate that type of error.

0% False Failures

		X	
		PASS	FAIL
Y	PASS		?
CUTOFF:____	FAIL	0	

0% False Passes

		X	
		PASS	FAIL
Y	PASS		0
CUTOFF:____	FAIL	?	

Compare the percentage of error in the remaining type of misclassification with the maximum acceptable level set in step number 1, and adjust the cutoff if necessary. Calculating the error rates for several alternative cutoff scores on Y should allow for making a reasonable choice.

6) If the information is available, determine the percentage of students who would be classified in the same category (pass or fail) on two successive testings using X. The percentage of classification errors using the final cutoff on Y should be the same or approximately the same as the percentage of students receiving a different classification when retested with X.

$X \longrightarrow Y$

Predicting Y From X

Introduction. In a few cases, it may be necessary to "equate" two instruments by predicting in only a single direction; i.e., using a linear or curvilinear regression approach. This approach has several problems: while minimizing the errors in predicting Y from X, it does not minimize errors in predicting X from Y. A conversion table of scores using this approach may be misleading if it is not distinguished from a conversion table that is two-directional. The regression approach is not truly "equating" because results are not symmetrical. The same equation that converts scores on X to scores on Y will not convert scores on Y to scores on X by solving for X (given Y).

If prediction rather than equating is truly the goal, the regression approach has the advantage of simplicity. Statistical packages are readily available and results are easily obtained. However, interpretation must be made with caution, as indicated previously.

Considerations. Sampling and administration procedures would be comparable to those discussed in the symmetric equating of X and Y. Possible regression solutions could include linear, quadratic, and cubic equations, as well as other nonlinear solutions. A comparison of the R^2 obtained from each of the equations should indicate which equation results in the most accurate prediction of Y.

Because the regression technique is not truly a test-equating procedure, more detail is not provided here. However, Angoff (1971) does suggest a linear equating method that is fairly simple to use, and the evaluator considering a regression approach to measurement may wish to consider this linear equating approach instead. The advantages in ease of interpretation may outweigh the slight disadvantage of mastering a relatively simple, new technique.

$X \longleftrightarrow Y$

Symmetric Equating of X and Y

Introduction. In developing a symmetric equating procedure that encompasses the full range of scores on X and Y, the evaluator or researcher attempts to derive an equivalent or at least comparable score on X for every score on Y, and vice versa. The direction of the conversion (from X to Y, or from Y to X) does not affect the results.

Considerations. Angoff (1971) suggests that the best way of ensuring equivalent scores is to use the equipercentile method of equating.[a] However, when the distributions of X and Y are similar, a linear alternative procedure is also suggested that may be considered an approximation of the equipercentile method.

Because the equipercentile method is so cumbersome, the evaluator choosing an equating procedure must consider how similar the distributions of X and Y actually are, and to what extent an approximation might be appropriate. Jaeger (1980) has provided a useful comparison of linear versus equipercentile methods of equating and mentions that differences in results between the two techniques are more noticeable at the extremes of score distributions, a crucial consideration for some types of testing. In addition, Jaeger also provides some guidelines and indices for choosing a test equating method for those persons considering a linear procedure.

Traditionally, the equipercentile method has been the method of choice,

[a] Due to the theoretical complexity and general unavailability of software, latent-trait models of equating are not considered here. Kolen (1980) suggests that equipercentile methods are still the most viable procedures for equating tests of differing difficulties, which is an issue arising in most cases of test equating.

and it will be the primary method discussed here. However, as Jaeger (1980) points out, with the increasing needs for multiple forms of the same test, more efficient methods may be needed in the future, and the reader who is interested in the simplicity of linear equating is referred to Jaeger (1980) and Angoff (1971, pages 568-571).

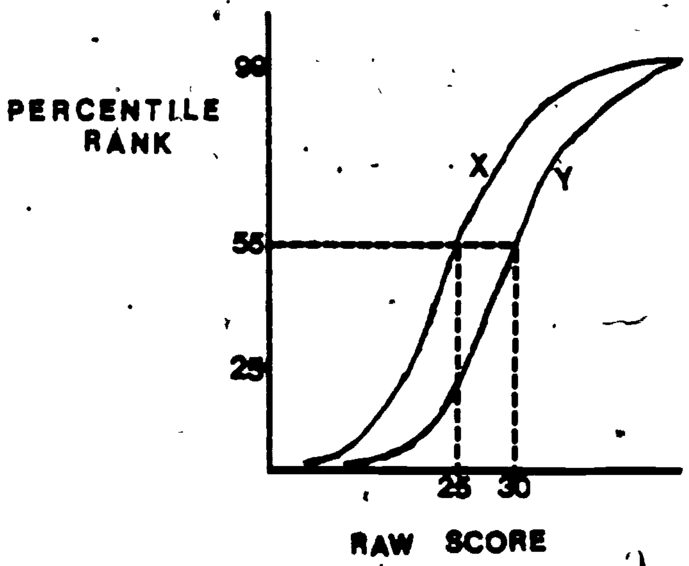
Suggested Steps

1) Sampling: In choosing a sample, there are several major factors to consider. Because the intent of this type of procedure is to equate along the full range of scores on X and Y, it is important that subjects in the study demonstrate the full range of abilities measured by the two instruments. The sample should reflect the ethnic and gender proportions of the population in the district as a whole. A score conversion table derived in this way assumes that both instruments are administered to a single group of individuals. However, a separate sample for each test may be an acceptable alternative if both samples are: a) large, b) drawn from the same population, and c) truly random.

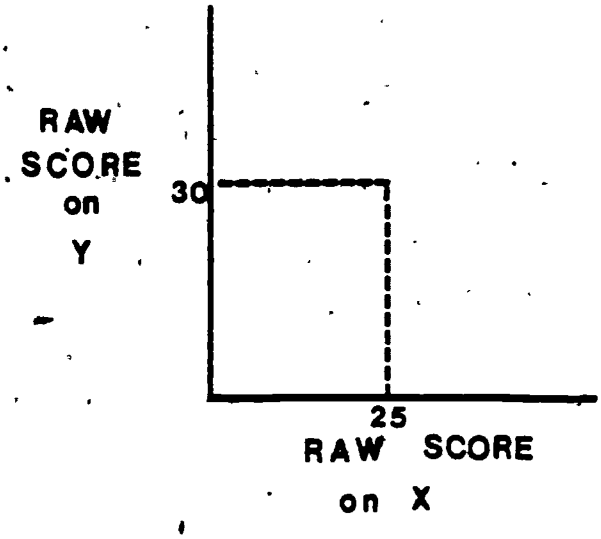
2) Administration: Ideally the entire sample would receive both instruments,with the order of administration random or counterbalanced. If the order of administration cannot be counterbalanced, administering the shorter test first (if the tests are of unequal length) should help to reduce fatigue effects. Depending on the length of the tests, at least one day to two weeks should elapse between administrations to minimize fatigue and practice effects as much as possible. (Too long between administrations may result in attrition of the sample and confounding maturational effects, especially if the order of administration is not counterbalanced.)

3) The steps in analysis are outlined in more detail by Angoff (1971). Briefly, midpercentile ranks or relative cumulative frequencies (the percentage of cases falling at or below each interval) are computed for each of the two distributions (X and Y).

4) The raw scores on X and Y are then plotted against the percentile rank.



5) The raw score to raw score conversion, based on the percentile ranks is then plotted. Angoff (1971) discusses methods of smoothing irregularities in these data, if needed.



Summary

It would be impossible to summarize in a single paper all of the theory and research that has been done in the area of test equating. There are many technical references that are both thorough and informative which persons with a serious interest in this area will want to read. However, this paper has attempted to summarize many of the practical issues facing the evaluator involved in test equating and also to provide some simple guidelines for such an endeavor.

References

Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed.) Washington, D.C.: American Council on Education, 1971, 508-600.

Baenen, N. and Curtis, J. Spring, 1980 TABS results for fifth and ninth graders--Technical Report. Austin: Office of Research and Evaluation, Austin Independent School District, 1980. Publication number: 79.40.

Guilford, J.P. Fundamental statistics in psychology and education. New York: McGraw Hill Book Company, 1965.

Jaeger, R.M. Some exploratory indices for selection of a test equating method. Paper presented to the American Educational Research Association, Boston, 1980.

Kolen, M.F. Comparison of traditional and latent-trait theory methods for equating tests. Paper presented to the American Educational Research Association, Boston, 1980.

Ligon, G. and Matter, M.K. Final technical report--equating study: 1970 California Achievement Tests, 1978 Iowa Tests of Basic Skills. Austin: Office of Research and Evaluation, Austin Independent School District, 1980. Publication number: 79.53.

Ligon, G. and Matusek, P. Equating the Comprehensive English Language Test to the Bilingual Syntax Measure II for identifying LESA students. Austin: Office of Research and Evaluation, Austin Independent School District, 1978. Publication number: 78.71.

Matusek, P. and Ligon, G. Individual versus group testing for identifying limited-English speaking ability students--an equating study of the Comprehensive English Language Test and Bilingual Syntax Measure. Paper presented to the American Education Research Association, Boston, 1980.