DOCUMENT RESUME

ED 395 038 TM 025 058

AUTHOR	Buras, Avery
TITLE	Test Equating Procedures: A Primer on the Logic and Applications of Test Equating.
PUB DATE	25 Jan 96
NOTE	19p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, January 25-27, 1996).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Equated Scores; *Item Response Theory; Models; *Raw Scores; *Statistical Analysis; Statistical Distributions
IDENTIFIERS	*Equipercentile Equating; *Linear Equating Method; Parallel Test Forms; Percentile Ranks

ABSTRACT

The logic and uses of test equating are discussed, including three methods of test equating. The focus is on the conceptual underpinnings of each test equating method, rather than on the mathematics of the procedures. Additional consideration is given to the assumptions of each method and its respective strengths and weaknesses. A commonly accepted definition of equivalent scores is based on the concept of equipercentile equating. The first step is to determine the percentile ranks of the distribution of scores on both instruments. The percentile ranks are then plotted against the raw scores for each instrument, and once they are graphed, equivalent scores can be obtained from the graph. The linear equating method is based on the assumption that the two forms of the test, which are designed to be parallel, will have essentially the same raw-score distributions, apart from minor differences in the mean and standard deviations. Item response theory (IRT), or latent trait theory, is an attempt to measure a person's performance on a test item as a function of the difficulty of the item and the examinee's performance on some unobserved, or latent, trait. The three IRT models are discussed, and conditions under which all equating methods are similar are reviewed. (Contains 1 table, 2 figures, and 18 references.) (Author/SLD)

Test Equating Procedures:

A Primer on the Logic and Applications of Test Equating.

Avery Buras

Texas A & M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA, January 25, 1996.

Abstract

The present paper will discuss the logic and uses of test equating. Three methods of test equating: “equipercentile equating;” “linear equating;” and equating with the one parameter or Rasch model of “item response theory” wili be included in the discussion. Rather than on the mathematics of the procedures, the focus will be on the conceptual underpinnings of each test equating method. Additional consideration will be given to the assumptions of each method and their respective strengths and weaknesses.

Introduction

During the semester break of my first year in college I had the opportunity to visit Toronto, Canada. Because I am originally from the Southern United States, my experiences dealing with snowy Winter conditions and cold temperatures were limited to pictures on post cards and viewing the Winter Olympics. Not only did I not have a concept of how cold it could be in Toronto in January, I had not bothered to learn my metric conversions while in the 6th grade. Consequently, I had no way of knowing how cold it was since I could not convert from Fahrenheit to Celsius and visa versa. Since the people I visited lived near the border of the U.S. and Canada, they needed a formula to constantly convert Fahrenheit to Celsius in order to understand such things as news casts and newspaper articles which were readily available from both countries. Their simple conversion was to multiply the degrees Celsius by 2 and then add 32 degrees to convert to Fahrenheit. For example, if it were 10 degrees Celsius, it could be readily calculated to 52 degrees Fahrenheit.

$$(10 \text{ X } 2) + 32^\circ = 52^\circ \text{ (Fahrenheit)}$$

Likewise, the conversion from 52 degrees Fahrenheit into Celsius could be accomplished by subtracting 32 from 52 and then dividing by 2.

$$(52^\circ - 32^\circ) \div 2 = 10^\circ \text{ (Fahrenheit)}$$

A more exact formula for converting the two measures of temperature include multiplication by 9/5 rather than 2 and 5/9 rather than division by 2. However, their simple formula allowed for quick conversions back and forth. It allowed for the understanding of one metric in terms of another metric. Unbeknownst to me at that time, was that their simple formula contained elements of "linear equating," one of the several methods of test equating to be covered later in the present paper.

Overview

In psychological and educational testing situations there are occasions when examinees are measured with two different instruments which are designed to measure the same psychological construct or the same content domain (Crocker & Algina, 1986). For example, a university professor teaching a large class may want to administer a test in the smaller lab sections of the course. Since the lab classes are spread-out throughout the week, the professor feels the need to have several parallel forms of the test. If the scores are not equally distributed the professor may need to establish equivalent forms on the two measures. The establishment of equivalent scores on varying forms of a test is called "horizontal equating" (Skaggs & Lissetz, 1986b).

Angoff (1971) reported that what is being sought in test equating is a conversion from the units of one form of a test to the units of another form of the same test. This notion carries with it two restrictions. The first restriction refers to the idea that the two instruments in question should be measuring the same characteristic, e.g., temperature, anxiety. Angoff noted (1971), "it makes no sense to ask for a conversion from, say, grams to degrees of Fahrenheit or from inches to pounds" (p. 564). While it may be mathematically possible to develop a linear function or a regression equation to predict grams from Fahrenheit, prediction is not the purpose of equating. Transforming units of the same construct is the purpose of equating (Angoff, 1971). The second restriction Angoff (1971) noted was that in order for two measures to be truly transformed, the resulting conversion should be independent of the individuals from whom the data were drawn to develop the conversion and should be applicable in future situations.

Equipercentile Equating

A commonly accepted definition of equivalent scores is base ' upon the concept of equipercentile equating. The equipercentile method originates in the definition of equivalent scores originally noted by Lord (1950) and Flanagan (1951), cited by Angoff (1971):

> 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 degree of reliability), may be considered equivalent if their corresponding percentile ranks in any given group are equal. (p. 563)

The first step in equipercentile equating is to determine the percentile ranks of the distribution of the scores on both instruments. The percentile ranks are then plotted against the raw scores for each instrument (Crocker & Algina, 1986). Figure 1 illustrates a plot of percentile ranks for two 20 item instruments. Once the percentile rank-raw score plots are graphed, equivalent scores can then be obtained from the graph by comparing the paired values of each instrument. For example, a score of 10 on form X is equivalent to a score of 9 on form Y and a score of 15 on form X is equivalent to a score of 11 on form Y. These paired values are then plotted onto one graph. Generally, the graph of the two percentile ranks of the two instruments is then smoothed analytically (Crocker & Algina, 1986).

Insert Figure 1 and 2 about here.

Linear Equating

The linear equating method is based upon the assumption that the two forms of the test which are designed to be parallel, will have essentially the same raw-score distributions, apart from minor differences in the mean and standard deviations (Angoff, 1971; Crocker & Algina, 1986; Jaeger, 1981;). When this assumption has been violated, equipercentile equating has been shown to yield better results than both linear equating and IRT models (Jaeger, 1981; Skaggs & Lissitz, 1986a). Nevertheless, when this assumption has been met, it should be possible to convert scores on one form of the measure into the same metric as the other form by employing a linear function. The function relating scores on Form X to scores on Form Y using linear equating is generally expressed as $Y^* = a(X-c) + d$; where "a" refers to the ratio of the standard deviations of Form Y over the standard deviation of Form X, "c" refers to the mean of form X, and "d" refers to the mean of form Y (Crocker & Algina, 1986).

This function is analogous to an equation that is derived in regression. The multiplicative constant of "a" is analogous to a B weight in regression and "c" and "d" are analogous to an additive constant. The reader is encouraged to understand this similarity but to also realize that test equating is not about prediction but is rather about calibration i.e., the re-expression of one measure in terms of another (Angoff, 1971).

There are three basic designs which are used to collect data in order to carry out a test equating project. In the first design, a large group of examinees are selected who are sufficiently heterogeneous in order to adequately sample all levels the of scores on both Form X and Form Y (Angoff, 1971). This large group is then divided randomly into two groups. Each group takes a different form of the test. The mean and standard deviations of both groups are calculated and can then be inserted into the linear function discussed above. Tablc 1 illustrates a linear equating example using design 1 with a small sample of 20 subjects, 10 who were administered form X and 10 who were administered form Y. The symbol, Y*, denotes the predicted scores on Form Y for those who have taken Form X. Since the standard deviations and means of both Forms were so similar very little adjustment is needed. Nevertheless, a raw score of 8 on Form X can now be expressed as a score of 8.3 on Form Y.

Insert Table 1 about here.

Design 2 is generally used when the test administrator has the luxury of more time to administer tests. In this design, both Form X and Form Y are administered to all subjects. In order to control for an order effect, half of the subjects receive Form X first and the other half of the subjects receive form Y first. The calculation of the multiplicative constant, i.e. "a" and the additive constants, i.e. "c" and "d" are slightly more involved with Design 2 (Angoff, 1971; Crocker & Algina, 1986) and are not included in the present paper.

Design 3 involves two randomly assigned groups each taking a different test and a common equating test administered to both groups. This common test is used as an "anchor test" (Parshall, Du Bose Houghton, & Kromrey, 1995). The calculation of the multiplicative constant, i.e. "a" and the additive constants, i.e. "c" and "d" are also more involved with Design 3 and are not included in the present paper. Crocker and Algina (1986) noted that Design 3 involves an additional assumption besides equal raw score distributions:

1. The slope, intercept, and standard error of estimate for the regression of X on Z in subpopulation 1 are equal to the slope, intercept, and standard error of estimate for the regression of X on Z in the total population.

2. The slope, intercept, and standard error of estimate for the regression of Y on Z in subpopulation 2 are equal to the slope, intercept, and standard error of estimate for the regression of Y on Z in the total population. (p. 460)

This design, also called an anchor test design, is the one most commonly used in test equating situations (Skaggs & Lissitz, 1986b). The anchor test may be internal, i.e., part of the content of both test, or the anchor test may be an external test as described above. For this design, intact or non-random groups may be used. The anchor test is designed to adjust for any between group differences that may be present (Parshall, Du Bose Houghton, & Kromrey, 1995).

Equating Using Item Response Theory

Item response theory (IRT), or latent trait theory, is an attempt to measure a person's performance on a test item as a function of the difficulty of the item and the examinee's performance on some unobserved, or latent trait (Skaggs & Lissitz, 1986b). The IRT model posits the relationship between a latent trait and the performance on a tcst designed to measure this latent trait. There are three IRT models. The first model is called the one-parameter or Rasch model. In the Rasch model, the only parameter that is dealt with is item difficulty. The second model, call the two-parameter model is essentially the same as the first with the addition of an item discrimination parameter. The third model includes both the difficulty and discrimination parameters and a parameter for guessing, and is called the three-parameter model. While there is extensive research on the two and three parameter models, the present paper will limit discussion to only the "Rasch" model.

IRT has at it's heart the notion of "Parameter Invariance" (Crocker & Algina, 1986; Divgi, 1986; Skaggs & Lissitz, 1986b). Basically, the item statistics that are derived from the application of an IRT model are independent of the sample of examinees to which the test was administered, sometimes referred to as "sample-free measurement." Also, the personal statistics, i.e., ability levels, obtained for examinees are independent of the sample of the items included on the test. This is sometimes referred to as "test-free measurement." But simply, IRT allows item independence and person independence (Crocker & Algina, 1986). Skaggs and Lissitz (1986b) stated,

> The basic advantage of IRT methods is that if data fit the model reasonably well, it is possible to demonstrate the invariance of item and ability parameter. That is, for a set of calibrated items, an examinee will be expected to obtain the same ability estimate from any subset of items. Moreover, for any subsample of examinees, item parameters will be the same. (p. 503)

For example, an examiner has a four item test with item 1 being the easiest and item 4 being the hardest, i.e., item 1 has a lower item difficulty than item 2. The examiner can give items 1 and 2 to person A and items 3 and 4 to person B. If person A answers item 1 correctly and misses item 2 and person B answers items 3 and 4 correctly, the examiner could then conclude that the latent ability level of person B was higher than the latent ability of person A (Crocker & Algina, 1986). This example illustrates the power of latent trait theory, if in fact, the item difficulty parameters do not vary from sample to sample. A more precise comparison between the two examinees could be accomplished with a larger sample of test items with varying difficulty levels (Crocker & Algina, 1986).

This discussion of the basics of IRT has been intentionally non-technical and lacking of sufficient details. It is well beyond the scope of this paper to have a more detailed discussion of IRT. The reader is referred to several sources for further reading on this topic (Hambleton & Swaminathan, 1985; McKinley & Mills, 1989; Skaggs & Lissitz 1986b).

In IRT applications, it still may be necessary to place two or more sets of test results into a common metric. Under IRT, test equating involves finding the coefficients of a linear transformation of the metric of one test to the metric of another (Baker, 1993). The basic function for IRT is given by (Baker, 1993):

$$\theta_i{}^* = A\theta_i + K,$$

where: A is the slope, K is the intercept, θ_i is the examinee's trait level parameter in the metric of the current test, and $\theta_i{}^*$ is θ_i expressed in the target test metric.

Algina and Crocker (1986) described a test equating method using a one parameter IRT model in which design 3 is used. Essentially, Form Y is given to group A and Form X is given to group B, with both forms containing a number of anchor items. Estimates of the item difficulties are derived. If the anchor items have different difficulty levels, an average of the differences in the item difficulties are obtained. This average differences is added to the appropriate test form and the scores on both tests can then be compared.

Evaluation of the Three Methods

As mentioned earlier, equipercentile equating is used when the assumptions of linear equating have not been met. However, the equating error is much larger with equipercentile equating (Angoff, 1982). Crocker and Algina (1986) suggest the use of linear equating procedures when the distributions of *z*- scores for X and Y are not too different.

With regard to the varying designs of linear equating, because design 2 includes information on all persons on both tests, it has a smaller standard error than does design 1 (Angoff, 1971). However, design 2 does require more time. Time may also be a factor in choosing design A over C as well. Like design 2, design 3 has a smaller standard error and is preferred when time is not an issue (Crocker & Algina, 1986).

Traditionally, equating methods are employed when there are several thousand examinees (Parshall, Du Bose Houghton, & Kromrey, 1995). However, some studies have been conducted to test the efficacy of using smaller samples and whether or not small samples have drastic effects upon the standard errors of the equating models for these small samples. Marks and Lindsay (1972) concluded that sample size directly increased test equating error and recommended that a minimal sample size be no lower than 250. Kolen (1985) examined the effects of sample sizes of 100 and 250. Samples of 250 appeared to have adequate accuracy of the estimates of standard errors of equating.

Since design 2 appears to be on of the most used designs across methods, research has examined the length of the anchor test, the extent to which the anchor test is similar to the content of the total test and the strength of the correlation between scores on the anchor test and scores on the total test. Budescu (1985) concluded that the best results were obtained when one half of the total test consisted of anchor items. Research has also indicated that the magnitude of the correlation between the total test and the anchor items is the single most important predictor of equating efficiency (Budescu, 1985; Parshall, Du Bose Houghton & Kromrey; 1995).

With regard to the one parameter or Rasch model, it appears that since guessing is such a distinct possibility i multiple choice tests, this model does not hold up well to multiple choice situations. Essentially, latent trait theory is based upon the assumption of unidimensionality. Unidimensionality refers to the idea that the only trait being measured is the so called latent trait. Guessing introduces another dimension or trait, since guessing ability also varies across people. Nevertheless, guessing is and can be addressed in the three parameter model.

Summary and Conclusion

In a review of the literature on test equating, Skaggs and Lissitz (1986a) summarized research in the area of horizontal test equating with the following 6 general statements:

> 1. No single method is consistently superior to the others. However, IRT equating seems to produce better results than conventional methods when the examinee samples taking each test form to be equated have not been randomly selected.
>
> 2. If the data are reasonably reliable, tests are nearly equal in difficulty, and samples are nearly equal in ability, then most linear, equipercentiles, and IRT methods will achieve satisfactory result.
>
> 3. The Rasch model is not robust to its assumption that no guessing has occurred in answering multiple-choice items. (p. 523)
>
> 4. Three-parameter model equating is not accurate when small sample sizes (less than 1,000 responses per item) have been used to estimate item parameters.

5. For small sample sizes, it is not clear which method works best or how stable the results will be.

6. Changes in the examinee population can bias a series of equatings that take place over an extended period of time.

The present paper was designed to be an introduction to the basics of test equating procedures. Emphasis was placed upon the concepts underlying the methods of test equating rather than on the statistical details of each method. While equating may at first appear to look like prediction (regression), the reader was encouraged to comprehend the difference between regression and test equating, i.e., the transformation of one metric into the metric of another. Also, since there are many methods to equate tests, the reader is encouraged to choose the method that is best suited for their particular needs and the characteristics of their samples.

References

Angoff, W.H. (1971). Norms, scales, and equivalent scores. In R.L. Thorndike (Ed.). Educational measurement (2nd ed.). Washington, DC: American Council on Education.

Angoff, W.H. (1982). Uses of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.). Handbook of methods of detecting item bias. Baltimore: Johns Hopkins University Press.

Baker, F.B. (1993). Equating tests under the normal response model. Applied Psychological Measurement, 17(3), 239-251.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. Journal of Educational Measurement, 22, 13-20.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Harcourt Brace Jovanovich, Inc.

Divgi, D.R. (1986). Does the Rasch model really work for multiple choice items? Journal of Educational Measurement, 23(4), 283-298.

Flanagan, J.C. Units, scores and norms. In E.F. Lindquist (Ed.), Educational Measurement. Washington, DC: American Council of Education, 1951, 695-763.

Hambleton, R.K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff.

Jeager, R.M. (1981). Some exploratory indices for selection of a test equating method. Journal of Educational Measurement, 18(1), 23-38.

Kolen, L.W. (1985). Standard errors of Tucker equating. Applied Psychological Measurement, 9, 209-223.

MacCann, R.G. (1989). A comparison of two observed-score equating methods that assume equal reliable, congeneric tests. Applied Psychological Measurement, 13(3), 263-276.

MacCann, R.G. (1990). Derivations of observed-score equating methods that cater to populations differing in ability. Journal of Educational Statistics, 15(2), 146-170.

McKinley, R.L. & Mills, C.N. (1989). Item response theory: Advances in achievement and attitude measurement. In B. Thompson (Ed.)., Advances in social sciences methodology (Vol. 1, pp. 71-135). Greenwich, CT: JAI Press..

Marks, E. & Lindsay, C. A. (1972). Some results relating to test equating under relaxed test form equivalence. Journal of Educational Measurement, 9(1), 45-56.

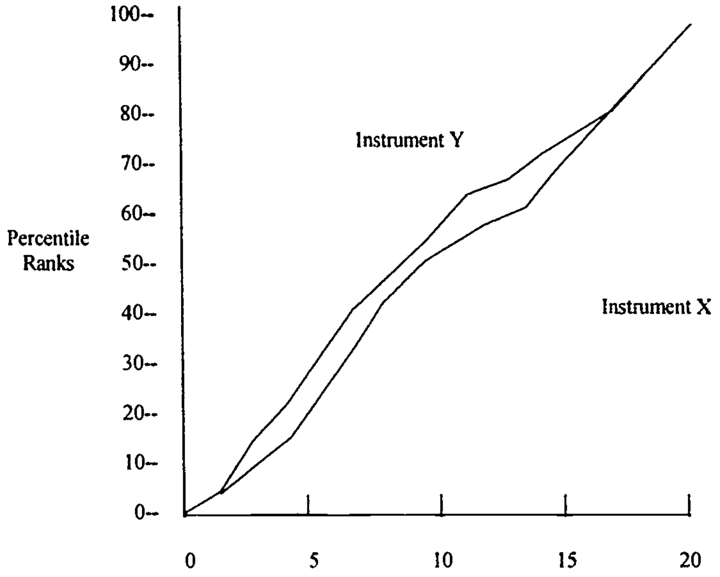
Lord, F.M. (1950). Notes on comparable scales for test scores (ETS RB 50-48). Princeton, NJ: Educational Testing Service.

Parshall, C.G., Houghton, P.D. & Kromrey, J.D. (1995). Equating error and statistical bias in small sample linear equating. Journal of Educational Measurement, 32(1), 37-54.

Skaggs, G. & Lissitz, R.W. (1986a). An exploration of the robustness of four test equating models. Applied Psychological Measurement, 10(3), 303-317.

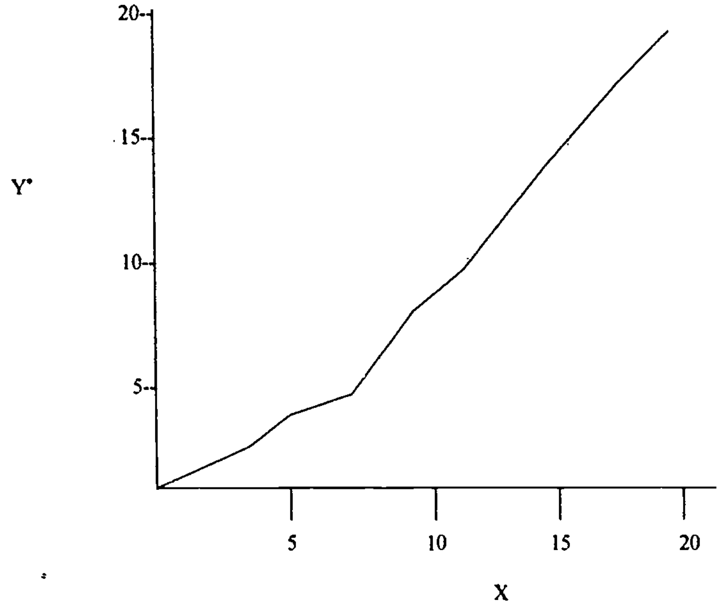
Skaggs, G. & Lissitz, R.W. (1986b). IRT test Equating: Relevant Issues and a review of recent research. Review of Educational Research, 56(4), 495-529.

Figure 1



Plot of Percentile Ranks of Two Instruemtms (Adopted From Crocker and Algina, 1986).

Figure 2



Plot of Percentile ranks of two 20 item intruements. (Adopted from Croker and Algina, 1986).

Table 1

	FORM X			FORM Y		FORM X	Y*
	5			9		5=	5.002732
	6			8		6=	6.106557
	4			4		4=	3.898907
	6			5		6=	6.106557
	7			6		7=	7.210382
	8			7		8=	8.314207
	9			8		9=	9.418032
	10			9		10=	10.52185
	7			10		7=	7.210382
	8			5		8=	8.314207
Sum X=	69		Sum Y=	71			
Mean X	6.9		Mean Y=	7.1			
SD X=	1.83		SD Y=	2.02			
			a=	1.103825			
			c=	6.9			
			d=	7.1			