DOCUMENT RESUME

ED 127 329 TM 005 414

AUTHOR	Rentz, R. Robert; Bashaw, W. L.
TITLE	Equating Reading Tests With the Rasch Model. Management Report.
INSTITUTION	Georgia Univ., Athens. Educational Research Lab.
SPONS AGENCY	National Center for Education Statistics (DHEW), Washington, D.C.
PUB DATE	Mar 76
CONTRACT	OEC-0-72-5237
NOTE	15p.; For related documents, see TM 005 415 and 416
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	Elementary Education; *Equated Scores; *Item Analysis; *Mathematical Models; *Probability; Reading Comprehension; *Reading Tests; Standardized Tests; Statistical Analysis; Vocabulary
IDENTIFIERS	*Anchor Test Study; *Rasch Model

ABSTRACT

To determine if Rasch Model procedures have any utility for equating pre-existing tests, data from the equating phase of the Anchor Test Study (ATS) were reanalyzed. This management report summarizes the work completed in the project, describes the differences in this project and that of the ATS, and presents recommendations and conclusions to the U.S. Office of Education. The tests involved included seven reading test batteries, each having one to three levels and two forms, and each having a vocabulary and comprehension subtest. There were 28 form-level combinations possible. Therefore, the concern was the simultaneous equating of 28 tests for each of vocabulary, comprehension, and total scores. The obvious difference between this project and the ATS is that this project used Rasch Model test calibrations and equating methods while the ATS used a variety of equipercentile and linear model methods. Other important differences are outlined. Finally, some conclusions and recommendations relative to any future equating effort that might be undertaken are made. Topics discussed include technical issues as well as design and cost considerations. (RC)

EQUATING READING TESTS WITH THE RASCH MODEL

MANAGEMENT REPORT

R. Robert Rentz

W. L. Bashaw

Educational Research Laboratory
College of Education
University of Georgia
Athens, Georgia

March, 1976

This report was prepared under Contract No. OEC-0-72-5237 by the Educational Research Laboratory of the University of Georgia for the National Center for Education Statistics, Division of Education, U. S. Department of Health, Education, and Welfare. The contractor was encouraged to exercise professional judgment in its contents; therefore, this report does not necessarily reflect positions and policies of the Government.

Introduction

The purpose of the project was to determine if Rasch Model procedures have any utility for equating pre-existing tests. Specifically, we reanalyzed the data from the equating phase of the Anchor Test Study.

The purpose of this management report is to summarize the work completed in the project, to describe the differences in our work and that of Educational Testing Service in the Anchor Test Project, and to present our recommendations and conclusions to the U. S. Office of Education.

Summary of Rasch Project

Data Organization

Tests involved included seven reading test batteries, each having from one to three levels and two forms, and each having a vocabulary and comprehension subtest. There were 28 form-level combinations possible. Therefore, we were concerned with the simultaneous equating of 28 tests for each of vocabulary, comprehension, and total scores..

We equated without regard to grade level of subjects, i.e., data from children who took a particular test were not subdivided by grade level when these children were members of more than one grade.

Theoretical Orientation

We reviewed both Rasch theory and equating procedures literature.

The following general principles evolved from this review.

1. In situations in which two tests reasonably conform to Rasch Model conditions and these two tests are administered to a single group of subjects; then equating simplifies to the determination of a single additive constant.

2. The stability of equating depends entirely on the stability of the raw score to ability score calibration; therefore, the observation of reasonably stable calibration implies equating stability.

3. Equated raw scores can be defined as scores corresponding to the same ability level, a definition that is analogous to equipercentile and linear model definitions.

Equating Errors

We developed equations for estimating standard errors of equating constants and also estimated these values directly from studies of calibration stability. We conclude that the major source of error is the usual measurement error. The error in the equating constants is trivial. There is an error involved in assigning raw scores to equivalent raw scores that can be avoided by using reference scales instead of raw score equating. The fourth possible error source is due to calibration instability which can be studied prior to undertaking equating studies..

Data-Model Conformity

Problems of assessing "model-data fit" were discussed. The most reasonable recommendation is that "fit" be determined by the degree to which specific objectivity is observed. That is, if various samples yield similar results, then invariance with regard to samples is present, which, in turn, implies adequate fit for equating purposes.

Such studies were conducted by comparing multiple random samples of size 500, 1000, 2000, and 4000; by comparing results of calibrations over all occurences of a test in the sampling design; by studying the differences in calibrations for racial and intelligence level subsamples for all tests; and by additionally studying STEP calibrations for subsamples divided by sex, grade, school system size, and school percentage of welfare families.

Our general conclusion from these studies was that results were adequately stable to support the use of the Rasch Model in equating these tests.

Methodology of Multiple Test Equating

Our procedures for developing equating constants and their standard errors were presented in detail. The specific methodology is easily modified for other possible sampling designs.

Methodology was also presented for using reference scales and for using user-developed new tests composed of any items on any of the tests included in our analyses.

Equating Tables

We present equating tables for both vocabulary and comprehension that allow a user to determine for a particular primary test form an equated score corresponding to any raw score obtained on any of the other 13 primary forms or on the appropriate secondary (parallel) form.

Due to the importance of assignment error, we present all possible assignment errors. Moreover, we attempt to solve the problem of assignment error by recommending and providing a reference scale (our National Reference Scale) for interpretting all obtained test scores on a common scale.

Test Calibration Data

For each of the 28 tests separately for vocabulary, comprehension, and total scores we present for each possible raw score the percentage of children earning that score, the score's corresponding ability estimate (unadjusted), the standard error of measurement for that ability, our National Reference Scale score, and the NRS score standard error. Also presented is the total test Kuder-Richardson formula 20 reliability estimate and the test's equating constant.

Item Analysis Data

For each item of each of the 28 tests, separately for vocabulary comprehension, and total scores, we present the following item data: difficulty (percentage correct), log easiness estimate, the corresponding standard error of the easiness estimate, the point-biserial of the item with ability estimates, the item characteristic curve slope, and an item mean square fit index.

We present for each of the 28 tests for each vocabulary, comprehension, and total scores summary data on all of the various item indexes. The summaries include frequency distributions, means, standard deviations, skewness indexes, kurtosis indexes, medians, and semi-interquartile ranges.

In addition, the relationship of these difficulties, easinesses, point-biserials, and slopes to the item mean square fit indexes is displayed graphically for all tests.

Rasch Project - Anchor Test Project Differences

The major objective of our project was to re-equate Anchor Test Project data using techniques of Rasch Theory. Thus, the obvious difference in our work is that we used Rasch Model test calibrations and equating methods while the Anchor Test Project used a variety of equipercentile and linear model methods. However, there were other important methodological and output differences which will be outlined here.

Equating Raw Scores

Results from the Anchor Test Project were based on data divided by grade level. Thus, they developed equating on somewhat different data than we used, as we kept together all data on a specific test. They present equating tables separately for each grade and include in each table only the seven tests considered by test publishers as appropriate for that grade. Our tables allow a user to administer out-of-grade-level tests and equate the obtained score to an appropriate in-level test.

Moreover, the Anchor Test Project did not provide tables for equating primary to secondary forms. Our tables allow for the conversion of secondary forms into primary forms.

Reference Scales

The Anchor Test Project tables provide no way to avoid assignment errors or to equate scores across test levels. Our National Reference Scale solves both of these problems. With it, any of the 28 tests can be given to any child and the resulting score can be interpreted free of assignment error.

Specifically, the NRS allows considerable opportunity to evaluate reading programs that involve growth over reading levels. Thus, data over a several year period can be evaluated on a common metric, allowing the opportunity for growth to be revealed. Such scaling is a necessary prerequisite to assessing growth without resorting to grade equivalent scores and their accompanying technical weaknesses.

It would be extremely valuable to obtain data for extending the NRS downward to reading readiness levels and upward to junior high school levels. Moreover, the freedom to use any of 28 tests as essentially parallel forms of each other can be quite valuable in the evaluation of programs requiring periodic assessment.

Comparisons of Tables

Direct comparisons of equating tables of the two projects was presented for selected test pairs using subsamples of subjects who were administered both tests in each test pair in the same order of testing. For each raw score on the base test three equated scores were obtained: the recommended Anchor Test Project value, the recommended Rasch Project value, and a subgroup conditional mean. The reader can scan the tables to determine how similar the various results are. Several such tables are presented which differ in regard to model-data fit and grade level. In general, tables are quite similar. Often both projects yield the same equated value, many values differ by one or two points, and only rarely are values different by more than three points. These differences are small relative to standard errors of measurement.

A discussion of the difficulty of comparing results is also presented. There is no legitimate way to say which is "best". The definition of "best" will be largely dependent upon the theoretical orientation of the reader. At least, with a strong model, such as the Rasch model, one can gather information on whether or not the results should be used. The equipercentile method does not lend itself to such tests.

Conclusions and Recommendations

The following sections of this report contain some conclusions and recommendations relative to any future equating effort that might be undertaken. The topics discussed include technical issues as well as design and cost considerations.

Raw Score Equating

There are three ways to achieve comparability between the scores on two or more tests. The first is to construct parallel forms, a process that is quite rigorous, resulting in isomorphic test score scales which by definition are equal. This procedure can only be accomplished at the test construction level and is mentioned here only to complete the context for the discussion that follows. The other two methods are the ones we have been concerned with in this study. They are raw score-to-raw score equating and raw score-to-reference scale equating; we will call them raw score equating and reference scale equating, respectively.

In reference scale equating, each test to be equated has its score scale translated into a single reference scale whose units may or may not be similar to its raw score scale. Two examples of these are the CEEB's Scholastic Aptitude Test scale of 200 to 800 and our own National Reference Scale for reading (see Volume I, Final Report) with an effective score range of 144 to 263 (for the tests included here). In equipercentile equating the scale of percentile ranks is the reference scale, linear equating used a z-score scale and Rasch equating is based on the log ability scale. Regardless, then, of the specific method of equating, each procedure has at least an implied reference scale, and these reference scales have their own unique properties. For example, the percentile rank scale is a description of the performance of the calibrating sample; and, as such, would differ from sample to sample. On the other hand, the Rasch ability scale does not depend on the calibrating sample; its values are invariant with respect to calibration by differing samples. Thus, regardless of specific method, tests to be equated are in fact translated into a particular reference scale whether that is the end product or not.

An additional step is taken with raw score equating. What happens is that two raw scores are assigned to be equal when they correspond to the same reference scale score. Furthermore, when an equivalent raw score on one test must be assigned an equivalent score on another test even when the difference between reference scale scores is large, the result is what we called "assignment" error". The magnitude of this error exceeds all other equating errors by a significant amount, as we have illustrated in our final report (secion 5.1, particularly Table 5.1.1). This factor alone argues against raw score equating and, indeed, that is our recommendation.

Vertical Equating

Another advantage of reference scale equating is that it permits the definition of a test scale across several levels of a test battery. A common scale that spans several grade levels would permit the measurement and description of growth and change. Equating several levels of the test battery by means of a common reference scale is called vertical equating and it is an important capability. Our National Reference Scale accomplishes this for the tests used and covers grades 4-6. We believe that the measurement of growth was a serious omission in the original conception of the Anchor Test design and ought to be included in any future equating efforts.

Evaluation of the Equating Process

In spite of our efforts to arrive at something that might be called a "standard error of equating", or for that matter the efforts of the Anchor Test Study's authors, a solution remains elusive. It is seemingly a simple matter to compute "standard errors" based on replications or to compute some root-mean-squared deviation from expectation; however, to conclude from that procedure the superiority of one method over another focuses on only the "consistency" property of an estimation. It is perhap in this case more important to focus on the property of bias to which neither we nor the Anchor Test Study directed ourselves. The issue is like making claims for test reliability without dealing with validity.

Additional research needs to be done to find a satisfactory way to compare and evaluate equating methods. At the present time, the fact of the matter is that the Rasch Model procedure and the equipercentile procedure are not strictly comparable. These two methods, along with linear equating, are based on different definitions of an equated score. Perhaps each does a good job of equating under its own definition but it is inappropriate to compare methods that attempt to do different things. At the same time, it still seems to be quite a meaningful question to ask: "If a person scores 43 on test A what would his score be if he had taken test B?" There may indeed be several answers to that question or perhaps we need to reformulate the question before we can get a satisfactory answer. Additional research needs to be done before these answers will be clear.

Possible Designs and Required Sample Sizes

The size of the sample will probably be the single most important factor in determining cost of any future equating study; however, the particular design that might be used is intimately associated with the sample size question. In the case where it is impossible to administer all tests to be equated to all examinees it would seem that some sampling procedure like that used for the Anchor Test Study would be most feasible. Angoff (1971) discusses several designs for data collection and Brigman (1976) has specifically compared three designs similar to the Anchor Test Study design; we will rely on he work in making some observations about designs.

Brigman (1976) compared the "full matrix" design (where tests are administered in all possible pairs as was done in the ATS) with two reduced models called a "chain" design and a "vector" design. These three designs are illustrated below:

Figure 1. Three Designs for Equating

	T_1	T_2	T_3	T_4
T_1	x/x	x	x	x
T_2	x	x/x	x	x
T_3	x	x	x/x	x
T_4	x	x	x	x/x

	T_1	T_2	T_3	T_4
T_1	x/x	x	o	x
T_2	x	x/x	x	o
T_3	o	x	x/x	x
T_4	x	o	x	x/x

	T_1	T_2	T_3	T_4
T_1	x/x	x	o	o
T_2	x	x/x	x	x
T_3	o	x	x/x	o
T_4	o	x	o	x/x

The situations depicted call for the equating of 4 tests. The "x/x" means that a particular test is given along with its secondary form. In all instances the row index test is administered first; and, since the matrices are symmetrical, it is obvious that when ever a particular test pair is administered the designs call for 2 replications with order-of-administration balanced.

As pointed out earlier Rasch Model Equating depends entirely on the estimation of the "equating constant", the translation factor which when added to the ability seale of one test in a test-pair equalizes the origin of both tests in that pair. The estimation of this constant requires only an estimate of the difference in average item difficulty between the two tests and constitutes the amount of adjustment necessary. Brigman found no essential difference between equating constants estimated from each design.

The importance of this finding is that any of these designs could be chosen for purposes other than adequate estimation of the equating constant. For example, the chain design might be best for the vertical equating of different levels of a test battery since adjacent lev ls could be administered to the same group whereas nonadjacent levels would be inappropriate (granted of course that we elminate cells $T_1 - T_4$ and $T_4 - T_1$ from the design). On the other hand the vector design would be appropriate for an equating study like the Anchor Test Study since one test could be administered in combination with all others at considerable reduction in the number of cells for which data were collected, 12 in the case of the Full Matrix design and 6 for the Vector design (ignoring diagonals in the above example).

In her study Brigman also investigated sample size, using cell sizes of 125, 250, and 500. Again there was no differencc. Our cwn work indicates that samples of 500 rroduces sufficient stability but that stability did increase up tc about 1000 and then began to level off.. Our conclusion about a required sample size is based on a per cell size of 500 to 1000; 500 would be inadequate, beyond 1000 would be wasteful, and no one would believe 125.

Estimates of Cost

It is not possible to estimate dollar costs for any future equating efforts; however, it is possible to identify cost factors that will determine dollar amounts. Two factors appear to us to be important: (1) the amount of data that need to be collected and (2) the extent to which the contractor's data processing capability has been developed. Any project will have a core of personnel which should be relatively constant across projects; however, projects may vary in personnel due to the two factors mentioned above. The same is true for supplies, materials and operations. We believe that considerable savings might be realized by funding equating studies in phases and we would like to deal briefly with one of these.

Equating Prerequisites

We have stressed the point many times that equating with the Rasch Model is simple and straight forward provided there is an acceptable degree of model-data fit. Evaluation of model-data fit ought to be separated from actual equating and furthermore, the funding should be separated. Model-data fit is the central question whenever the Rasch Model is to be applied to existing tests. Studies of this sort could be made without collecting additional data if for example publishers could be persuaded to let a contractor use data they already have, an arrangement which we have found successful in the past. If fit studies prove successful, then there need be little concern for elaborate and costly sampling plans for an equating phase.