DOCUMENT RESUME

ED 190 605 TM 800 397

AUTHOR Kolen, Michael J.

TITLE Comparison of Traditional and Latent Trait Theory Methods for Equating Tests.

PUB DATE Apr 80

NOTE 24p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Achievement Tests: *Difficulty Level: *Equated Scores: Guessing (Tests): High Schools: Latent Trait Theory: Quantitative Tests: *Test Items: True Scores: Vocabulary Skills

IDENTIFIERS Equipercentile Equating: *Iowa Tests of Educational Development: Linear Equating Method

ABSTRACT

Results from equipercentile, linear, and latent trait equating of the vocabulary and quantitative thinking tests of the Iowa Tests of Educational Development were compared. The study entailed both the equating of forms (of similar difficulty) and the equating of levels (of differing difficulty). The goal was to equate seventh edition tests to those of the sixth edition. The data were item responses from a representative sample of 10,728 Iowa high school students. One-, two-, and three-parameter logistic latent trait methods were used. The results from the equating methods were compared using a cross-validation criterion which measured the closeness of converted score distributions to actual score distributions for randomly equivalent groups. The one-parameter methods results were judged inadequate for equating tests differing in difficulty, possibly because of prevalent examinee guessing. The three-parameter methods results were promising although two problems were discussed which require further study. Presently, equipercentile procedures may be the most viable for equating tests of differing difficulty. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190605

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

COMPARISON OF TRADITIONAL AND LATENT TRAIT THEORY METHODS FOR EQUATING TESTS*

Michael J. Kolen

Hofstra University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Kolen, Michael J.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the American Educational Research Association in Boston, April, 1980.

* The author greatly appreciates the assistance of Leonard S. Feldt and Douglas R. Whitney in all phases of the study. Helpful suggestions by Robert A. Forsyth, H. D. Hoover, William E. Coffman, and an anonymous AERA reviewer are also appreciated. Any inaccuracies are the sole responsibility of the author.

TM 800 397

ERIC

ABSTRACT

Results from equipercentile, linear, and latent trait equating of the vocabulary and quantitative thinking tests of the Iowa Tests of Educational Development were compared. Tests of similar and of differing difficulty were equated. The data were item responses from a respresentative sample of 10,728 Iowa high school students. One-,two-,and three-parameter logistic latent trait methods were used. The results from the equating methods were compared using a cross-validation criterion which measured the closeness of converted score distributions to actual score distributions for randomly equivalent groups.

The one-parameter methods results were judged inadequate for equating tests differing in difficulty, possibly because of prevalent examinee guessing. The three-parameter methods results were promising although two problems were discussed which require further study. Presently, equipercentile procedures may be the most viable for equating tests of differing difficulty.

COMPARISON OF TRADITIONAL AND LATENT TRAIT THEORY METHODS FOR EQUATING TESTS

Achievement test batteries are typically published in several parallel forms with different levels for different grades. The forms and levels of each test making up the battery must be equated to one another. That is, every score on a given form or level must be translatable into a score value on any other form or level of that test.

Equipercentile and linear methods traditionally have been used to equate tests (Angoff, 1971). Latent trait methods recently have been advocated as possible improvements over the traditional methods (Lord, 1977; Wright, 1977). Lord (1977) argued from theoretical considerations that traditional equating methods are not appropriate for equating tests of differing difficulty, whereas latent trait theory methods have the capacity to provide an appropriate equating in this case.

Lord's (1977) definition of equating implies that exact equating is possible only when the tests to be equated measure the same unidimensional ability. Achievement tests covering abilities encountered over a range of grades are probably not unidimensional. However, the usefulness of a score scale will be severely limited unless it spans all of the levels for which the battery is intended. Thus, equating of levels must be attempted even when unidimensionality does not hold.

The intent of this study was to compare the end results of two traditional and seven latent trait theory equating schemes using data from the 1978 equating project of the Iowa Tests of Educational Development (ITED). The study entailed both the equating of forms (of similar difficulty) and the equating of levels (of differing difficulty).

A cross-validation group was used to establish a criterion for comparing the results of the equating methods. A cross-validation summary statistic was calculated which was a measure of the closeness of converted score distributions for stratified randomly equivalent groups. The goal of the study was to identify the method or methods which were "best" according to this criterion and to examine idiosyncracies of the equating schemes for the equating of a two-level high school achievement test battery.

Test Equating Definitions

Non-parallel tests X and Y (that is, tests measuring the same unidimensional ability but differing in difficulty or reliability) can be considered to be equated if any two examinees of equal true ability, one taking test X and the other taking test Y, would be expected to obtain the same score when performance on test X and test Y are expressed on a common score scale. This will be referred to as the definition of equating for non-parallel tests.

According to Lord (1977, p.128) tests X and Y can be considered to be equated "... if and only if it is a matter of indifference to each examinee whether he is to take test X or test Y". This definition implies that the definition of equating non-parallel tests holds. It also implies that for any population of examinees with equal ability the distribution of observed scores on test X will be identical to that of test Y. Hence, the standard error of measurement (as well as the higher order moments) for any individual (or group of individuals of identical ability) must be the same for test X as for test Y when the scores are expressed on the common score scale. Lord (1977) explained that this definition can be expected to hold only when test X and test Y are carefully constructed parallel forms. Hence, the above definition will be referred to as the definition of equating for parallel tests. Note that both of the equating definitions require that the tests to be equated measure the same unidimensional ability.

For equipercentile or linear equating to be exact, the definition of equating for parallel tests must hold. This is necessary because these methods require that a common score scale be constructed such that identical expected frequency distributions for the two tests will result for any subgroup of examinees. Thus, conventional equipercentile or linear equating can be strictly used only with parallel tests. In theory, the latent trait equating methods proposed by Lord (1977) and Wright (1977) can be used to equate both parallel and non-parallel tests under the definitions discussed here.

Review of Equating Research

Two types of studies have been carried out which assess the adequacy of various equating methods. In the first type the adequacy of a single equating scheme is assessed by examining the similarity of the results obtained from disparate groups. The groups may differ in such characteristics as ability, socio-economic status, or race. These studies are based on the principle that, if the definition of equating for non-parallel tests holds, then equatings based on diverse groups should be identical, apart from sampling error. In the second type of study the end results of various methods have been compared to one another. The present study is of this latter type.

Studies Using Different Groups

Linn (1975, p.207) concluded that the equipercentile equating of elementary school reading tests of similar difficulty in the Anchor Test Study (Loret, Seder, Bianchini, and Vale,. 1974) was "quite satisfactory for most practical purposes." In a reanalysis, Slinde and Linn (1977) focused on equipercentile equating across grades. They concluded that when tests differed substantially in difficulty, equipercentile results were inadequate.

Several studies have examined the effects of using different sub-groups of Rasch scaled items (e.g. odd-even, easy-difficult) on Rasch ability estimates for differing groups of examinees . One set of studies (Curry, Bashaw, and Rentz, 1978; Tinsley and Dawis, 1975; Whitely and Dawis, 1974; and Wright, 1968) led to the conclusion that the ability parameter is, in fact, invariant over item subgroups.

In another set of studies (Loyd and Hoover, 1979; Slinde and Linn, 1978; and Slinde and Linn, 1979) Rasch-based equatings of tests of substantially different difficulty were found to be highly dependent on the ability of the group[1]. Gustafsson (1979b) and Slinde and Linn (1979) hypothesized that the effects of guessing may have contributed to differences in equating results. Gustafsson (1979a) showed that examinee guessing on a test may result in a negative correlation between item difficulty and discrimination. Since Slinde and Linn (1979) found evidence that such a negative correlation does occur, the effects of guessing may have been a factor in their results.

<u>Studies Comparing Methods</u>

Lord (1977), Marco (1977), and Woods and Wiley (1977, 1978) compared some conventional and latent trait theory equating methods. These studies indicate that the equating schemes studied produce somewhat different results.

Rentz and Bashaw (1977) reanalyzed The Anchor Test Data using Rasch equating procedures and concluded that the Rasch and equipercentile equating results were reasonably similar. However, Slinde and Linn (1977) pointed out that the equipercentile method was not adequate for tests of differing difficulties. Thus,

[1] It should be noted that while Rasch model equating procedures (Wright, 1977) were used in these studies, Rasch Model test construction procedures (Wright, 1977) were not.

it cannot be determined whether the Rasch procedures provided any additional benefits over "inadequate" equipercentile methods.

Marco, Petersen, and Stewart (1979) compared a variety of equipercentile, linear, and latent trait theory equating methods for equating the verbal portion of the Scholastic Aptitude Test. When a test was equated to itself with an anchor test of similar difficulty all but one of the methods appeared to be satisfactory. The exception was one of the variations of the equipercentile method. The linear equating procedures appeared to produce the most accurate equating in this situation.

When tests of different difficulty were equated, the latent trait methods were superior and the linear methods clearly inferior. However, Marco et al. (1979) noted that the criterion used for judging the superiority of equating methods may have been biased against certain of the methods for equating tests of differing difficulty. Hence, conclusions based on these results are very tentative.

Summary

The studies reviewed here indicate that traditional and latent trait methods can be expected to produce adequate equating results when parallel tests are equated. Little empirical evidence exists for the superiority of any equating method for tests of differing difficulty. It appears that linear equating is not a sound procedure. Problems have also been found with equipercentile and Rasch methods. If, as Slinde and Linn conclude, examinee guessing accounts for the failure of the Rasch method, then the three-parameter logistic model should provide a more suitable approach with tests of differing difficulty.

Equating Problem

The ITED

The seventh edition of the ITED includes separate tests in seven areas. The tests are designed for administration to high school students. A description of the tests and the philosophy underlying their construction is presented in the ITED Manual for Administrators and Testing Directors (1972). Only two of the ITED tests - vocabulary and quantitative thinking - were analyzed in the present study.

The sixth edition of the ITED consists of one level administered in all grades. The new seventh edition of the ITED has two levels, with one pair of parallel forms (X-7 and Y-7) at each level. Level I of the seventh edition is designed for administration to students in grades 9 and 10 and Level II for administration in grades 11 and 12.

The sixth edition vocabulary and quantitative thinking tests contain 40 and 36 items and have time limits of 15 and 45 minutes, respectively. Level I and Level II of the seventh edition forms each have the same number of items and time limits as their sixth edition counterparts. One-third of the seventh edition items are common to Level I and Level II. No items contained in the sixth edition are included in the seventh edition.

In general, Level I of the seventh edition tests are easier than their sixth edition counterparts. Level II of each test is similar in difficulty to the sixth edition version.

Equating Project for the Seventh Edition

The goal of the ITED equating project was to equate seventh edition tests to those of the sixth edition. The study was based on the scores of 10, 728 high school students from 34 Iowa schools. The schools chosen for inclusion in the project represented the full range of averages exhibited by Iowa schools, as inferred from their previous year's performance.

Within each 9th and 10th grade classroom included in the project, forms X-6, X-7 Level I, and Y-7 Level I of the entire battery were administered to random thirds of the students. Within each 11th and 12th grade classroom, forms X-6, X-7 Level II, and Y-7 Level II of the whole battery were administered to random thirds of the students. Because of the random assignment of forms to students within each classroom, the three groups at each level can be considered stratified random samples -- stratified with respect to class and school. Each pupil took only one form of the tests.

For the present study, students with missing scores, zero scores, or perfect scores were eliminated because latent abilities of such students cannot be estimated with latent trait estimation procedures. The number of 9th through 10th grade students included in the present study ranged from 1,883 taking Level I of form Y-7 of the vocabulary test to 1,925 taking form X-6 of the vocabulary test. Similarly, the numbers of 11th through 12th graders ranged from 1,579 taking Level II of form X-7 of the vocabulary test to 1,643 taking form X-6 of the quantitative thinking test. Every third student within each form and test combination was withheld from the equating portion of the study. Their scores were used as a cross-validation check for the equating. This aspect of the study will be explained in a later section of this paper.

Equating Methods

One equipercentile, one linear, and seven latent trait theory equating methods were compared. Angoff (1971) has provided a thorough discussion of linear and equipercentile methods. Overviews of latent trait theory and latent trait theory equating have been supplied by Baker (1977); Cook and Hambleton (1977); Hambleton, Swaminathan, Cook and Eignor (1979); Kolen (1979); and Lord (1975,1977). Overviews of the Rasch model, which is one of the latent trait models, have been provided by Wright (1977) and Wright and Stone (1979). The following discussion assumes familiarity with at least some of these references.

The X-6 raw score scale was used as the common score scale. For those equating methods requiring interpolation, linear interpolation was used as a time-saving device. Identical procedures were followed for forms X and Y of the vocabulary and quantitative thinking tests.

Equipercentile and Linear Methods

Method IA-1 described by Angoff (1971) was used for linear equating and Method IA-2 for equipercentile equating. First, Level I of each seventh edition test and form was equated to form X-6, using the combined data for grades 9 and 10. Then, Level II of the seventh edition was equated to form X-6 using only the 11th and 12th grade data.

Latent Trait Methods

One-, two-, and three-parameter logistic latent trait models were used. Additionally, a modified one-parameter model was included, in which the common slope of the item characteristic curves was allowed to differ from the sixth to the seventh edition forms. Similar procedures were followed for each of the latent trait models.

The ability and item parameters were estimated using the Wood, Wingersky, and Lord (1976) LOGIST computer program. Because one-third of the items were common to the two levels, the parameters for Levels I and II of each seventh edition test form were estimated using simultaneous procedures. The parameters for the sixth edition tests were estimated using standard LOGIST procedures.

The item and latent ability parameters for the seventh edition tests were then equated to the sixth edition scale. This was accomplished by using the fact that the randomly equivalent groups taking forms X-6, X-7 and Y-7 would be expected to have identical distributions of latent ability, apart from sampling error. For the one-parameter model, the mean latent ability was used to equate seventh edition ability and item parameter estimates to the sixth edition scale. For the three remaining latent trait models, linear equating was completed, using the mean and standard deviation of the latent ability estimates. Hence, forms X-6, X-7, and Y-7 were on the same latent ability scale for each of the four latent trait models.

Estimated true score equating. The estimated true score (Lord, 1977) of an individual with a given estimated latent ability is equal to the sum, over items, of the estimated probability of correctly answering each item. Using the non-linear estimation procedure ZBRENT (IMSL, 1978), edition seven estimated true score equivalents of sixth edition integer scores were found. Similarly, sixth edition estimated true score equivalents of seventh edition integer scores were found. Each test, form, and level combination of the seventh edition was equated to the corresponding sixth edition test using these procedures with the four latent trait models.

Estimated observed score equating. Lord (1975, 1977) has shown that after the latent trait parameters are estimated, an estimated observed distribution of raw scores can be constructed using the generating formula for the generalized binomial. Separate estimated observed score distributions were constructed for forms X-7 Level I, X-7 Level II, Y-7 Level I, and Y-7 Level II as well as for the 9th-10th and 11th-12th graders taking form X-6. Each form-level of edition seven was then equated to the edition six raw score scale using equipercentile equating of the appropriate estimated observed score distributions. This procedure was followed for the modified one-parameter, two-parameter, and three-parameter models.

Methods and Abbreviations

The nine equating methods and their abbreviations are: 1) Conventional equipercentile (EQUI); 2) Conventional linear (LIN); 3) Modified one-parameter estimated true score (TEQM1); 4) Modified one-parameter estimated observed score (ESTOBM1); 5) two-parameter estimated true score (TSEQ2); 6) two-parameter estimated observed score (ESTOBS2); 7) three-parameter estimated true score (TSEQ3); 8) three-parameter estimated observed score (ESTOBS3); 9) One-parameter estimated true score (TSEQ1)

Evaluation Procedures

No demonstrably superior criterion for judging the relative accuracy of the various equating methods was available in this study. Therefore, the primary evaluative technique was to estimate the stability of the results when applied to a new, independent sample.

Two frequency distributions of raw scores on the sixth edition were constructed for students in the cross-validation samples -- one for 9th-10th graders and another for 11th-12th grade students. Likewise, frequency distributions for the cross-validation sample students taking forms X-7 and Y-7 were constructed. Using the results from each equating scheme the X-7 and Y-7 scores were converted to the X-6 scale.

The cross-validation criterion was the mean (over examinees in the X-6 cross-validation distribution) squared difference between sixth edition integer scores and seventh edition converted (equated) scores with identical percentile ranks in randomly equivalent cross-validation distributions. Smaller values of this index reflect greater consistency between the sixth edition and converted seventh edition cross-validation distributions. For any particular test, form, and level combination of the seventh edition, smaller values of the index were interpreted as indicating more stable equating for that method.

Estimated true scores below the "pseudo-chance" level of a test are undefined for the three-parameter logistic model. In order to include the three-parameter estimated true score method in the cross-validation, scores of one on any pair of tests were arbitrarily considered to be equivalent; "missing" equivalents below the "pseudo-chance" level were arrived at by linear interpolation

Results

Cross-validation statistic values are shown in Table 1.

Insert Table 1 About Here

Based on the cross-validation statistic values, ranks were assigned to the methods for each test, form, and level combination. Within each level, the four test and form combinations were treated as randomly selected blocks and the nine equating methods as treatments in the calculation of two Friedman statistics (Conover, 1971). A Friedman statistic surpassing the appropriate critical value indicates that overall, the methods differed in the cross-validation. The Friedman statistic for Level I was 25.0 $(p<.01)$ for Level II 16.8 $(p<.05)$. Kendall's coefficient of concordance(Conover, 1971), a measure of the average correlation among ranks,was 0.78 for Level I and 0.52 for Level II.

At Level I of the tests the three-parameter estimated observed score distribution method appeared to produce the most accurate results. The equipercentile method produced more accurate results, at least for the quantitative thinking tests, than the remaining methods. The linear scheme produced the least accurate results and the one-parameter true score equivalents method the next least accurate equating results. The results from the other methods appeared to be indistinguishable at Level I.

For Level II, the three-parameter estimated true score equivalents scheme tended to produce the most accurate results. In all cases, the one-parameter estimated true score equivalents method produced mcre accurate results than the modified one-parameter estimated true score equivalents methods. The results from the other methods seemed to be indistinguishable.

Discussion

Equirercentile vs. One-Parameter Methods

One notable finding was that the equipercentile method produced more accurate cross-validation results than the one-parameter or modified one-parameter true score equivalents methods for Level I of both tests. Level I was a downward extension of the sixth edition test and, hence, was an easier test. Therefore, a combination of examinee guessing and the equating of tests of differing difficulty was present in the equating of Level I of the seventh edition to the edition six score scale.

Note that the one-parameter true score equivalents equating procedures are identical to Rasch equating procedures(Wright, 1977) and differ from Rasch model equating only in the procedure used in test construction. As Gustafsson (1979a, 1979b) and Slinde and Linn (1979) have pointed out, if guessing is prevalent with the Rasch model then item difficulty and discrimination could be expected to be negatively correlated. For the two-parameter logistic model, the correlations between item difficulty and discrimination parameter estimates for total tests ranged from -0.4517 to -0.7081 (Median - -.6813). The hree-parameter logistic model correlations ranged from 0.0340 to 0.3670 (Median = 0.1069). Thus, the inclusion of the lower asymptote parameter resulted in minimal correlations. These findings suggest that the failure of the one-parameter and modified one-parameter schemes to take guessing into account may have reduced their effectiveness at Level I of the tests. Since Level I of the seventh edition was of substantially lesser difficulty than the sixth edition tests, these data are consistent with the Gustafsson (1979a, 1979b) and Slinde and Linn (1979) conclusion that the prevalence of examinee guessing may have an adverse effect on the equating of test of differing difficulty using the one-parameter methods.

Another notable result was that the modified one-parameter true score equivalent (TEQM1) method produced more accurate cross-validation results for Level I and less accurate results for Level II of both seventh edition tests than did the one-parameter true score equivalents (TSEQ1) method. These one-parameter and modified one-parameter schemes differ only in the manner in which the overall (common) discrimination of the item characteristic curves is handled. For the one-parameter method, the common item characteristic curve discrimination for the seventh edition was forced to equal the common discrimination of the sixth edition curves. For the modified one-parameter scheme, while there was a common discrimination for all seventh edition curves of a particular form it was allowed to differ from the common discrimination of the sixth edition curves.

Negative correlations were found between the difficulty and discrimination parameter estimates of the two-parameter model. Hence, the average discrimination of Level I items, when guessing was not taken into account by a lower asymptote parameter, was greater than that of Level II. Comparing the cross-validation findings from the one-and modified one-parameter true score equivalents methods, it would appear that the items of Level II of the seventh edition had item discrimination similar to those of the sixth edition. The items of Level I probably had greater item discrimination than those of the sixth edition. The correlation between item difficulty and discrimination was probably a result of examinee guessing. Therefore, the differences between the cross-validation results for the one-and modified one-parameter true score equivalents may have resulted from differential effects of examinee guessing on Level I and Level II of the seventh edition.

Three-Parameter Logistic Model

The three-parameter estimated observed score distribution method tended to produce the most accurate cross-validation results at Level I of the tests but results of moderate accuracy at Level II. The three-parameter estimated true score equivalents method tended to produce the most accurate cross-validation results at Level II but results of moderate accuracy at Level I.

No convincing explanation for the three-parameter logistic model results could be found. However, two interesting facts may be noted. First, the three-parareter estimated true score equivalents method does not provide estimated true scores below the "pseudo-chance" level of the test, that is, below the sum of the lower asymptote parameter estimates. (Interpolation was used to arrive at equated scores below this level in the cross-validation analyses). Thus, the estimated true score scale is a condensed version of the observed score scale. This condensed scale probably differs from the observed score scale near the "pseudo-chance" level of the test and to a lesser extent along the entire score scale. Possibly, similar condensing of score scales occurs for tests of similar lifficulty but differential condensing occurs for tests of unequal difficulty. If so, this partially explains the finding that the three-parameter estimated true score equivalents method produced the most accurate cross-validation results at Level II and comparatively less accurate results at Level I.

Second, when the three-parameter model parameters are estimated, the LOGIST program may be weak in accurately assessing the lower asymptote parameter. In this case, the lower asymptote estimate of those items for which this difficulty exists are fixed at a common value. Of the seventh edition lower asymptote parameters estimated, 92% of the items on Level I only, 53% of the items common to both levels, and 39% of the items on Level II only, were fixed at common values. This possible failure for the items in Level I only probably had an effect on the equating; the precise effect is not clear, however.

Linear Method

The results make it clear that the linear method is not satisfactory when equating tests of unequal difficulty.The results for equating tests of similar difficulty suggest that the relationship between scores on the sixth and seventh edition tests are not linear throughout the entire range of scores.

Comments on Cross-Validation

The cross-validation criterion was designed to be a measure of stability over random sampling rather than a measure of accuracy of equating. The criterion was developed to indicate which, of a number of equating methods, produced the most consistant results in the equating of a set of pre-existing achievement tests rather than tests designed specifically to fit any one of the equating models. The exclusive use of comparisons, and the fact that the sampling distribution of the cross-validation statistic is unknown, precludes definitive statements about the consistency of the methods. Studies such as Slinde and Linn (1977, 1978, 1979) provide evidence of accuracy in a more absolute sense. Both comparative and absolute accuracy studies need to be completed.

Conclusion

The one-parameter models were found to produce inadequate results, perhaps, because of the prevalence of examinee guessing. Unless examinee guessing is eliminated from test performance, possibly by using the Wright and Stone (1979) procedures for discarding items showing a lack of fit to the Rasch model, the present research and that of Slinde and Linn (1979) suggest that inadequate results will occur when tests of differing difficulty are equated.

The three-parameter logistic model seems promising as a model for test equating. However, questions about the effects of condensing the score scale with the three-parameter estimated true score equivalents method and of the possible inadequate estimation of the lower asymptote parameter still need to be answered.

The equipercentile method produced reasonably adequate results. This method may presently be the most viable for equating tests that differ in difficulty to the extent that they differed in the present study, even though the equipercentile method could not be expected to produce a theoretically "perfect" equating in this case.

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.

Baker, F. B. Advances in item analysis. Review of Educational Research, 1977, 47, 151-178.

Cook, L. L. & Hambleton, R. K. application of latent trait models to the development of norm-referenced and criterion-referenced tests. Paper presented to National Council of Measurement in Education Toronto,1978.

Conover, W. S. Practical nonparametric statistics. New York: Wiley, 1971.

Curry, A. R., Bashaw, W. L, Rentz, R. R. Invariance of Rasch model ability parameter estimates over different collections of items. Paper presented to American Educational Research Association, Toronto, 1978.

Gustafsson, J.-E. Testing and obtaining fit to the Rasch model. Paper presented to American Educational Research Association, San Francisco, 1979a.

Gustafsson, J.-E. The Rasch model in vertical equating of tests: A critique of Slinde and Linn. Journal of Educational Measurement, 1978, 16, 153-158.

Hambleton, R. K. & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R.L, Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory, models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hieronymous, A. N. & Lindquist, E. F. Manual for administrators, supervisors, and counselors. Forms 5 & 6. Iowa Tests of Basic Skills. Iowa City, Ia.: Iowa Testing Programs, 1972.

IMSL Library 1, (Fortran IV) IBMS/370-360. 7th ed. Houston: International Mathematical and Statistical Libraries, Inc. 1978.

ITED Manual for Administrators and Testing Directors. Forms X-6 & Y-6. Iowa City, Ia.: Iowa Testing Programs, 1972.

Kolen, M. J. Comparisons of equipercentile, linear and selected latent trait methods for equating forms and levels of the seventh edition of the Iowa Tests of Educational Development. Unpublished Ph.D. Dissertation, University of Iowa, 1979 .

Linn, R. L. Anchor test study: The long and the short of it. Journal of Educational Measurement, 1975, 12, 201-214.

Lord, F. M. A survey of equating methods based on item characteristic theory, Research Bulletin 75-13. Princeton, N. J.: Educational Testing Service, 1975.

Lord,F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 177-138.

Loret, P. G., Seder, A., Bianchini, J. C. & Vale, C. Anchor test study final report: Project report and volumes 1 through 30. Berkeley, Calif.: Educational Testing Service, 1974.

Loyd, B. H. & Hoover, H. D. A comparison of methods of vertical equating. Paper presented to National Council on Measurement in Education, San Francisco, 1979.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977 14 139-160.

Marco, G. L., Petersen, N. S. & Stewart, E. E. A test of the adequacy curvilinear score equating models. Paper presented to 1979 Computer Adaptive Testing Conference, Minneapolis, 1979.

Rentz, R. R. & Bashaw, W. L. The national reference scale for reading: an application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-180.

Slinde, J. A. & Linn, R. L. Vertically equated tests: Fact or phantom? Journal of Educational Measurement, 1977, 14, 23-32.

Slinde, J. A. & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Slinde, J. A. & Linn, R. L. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement, 1979, 16, 159-165.

Tinsley, H. R. & Dawis, R. V. An investigation of the Rasch simple logistic model: Sample free item and test calibration, Educational and Psychological Measurement, 1975, 35, 325-339.

Whitely, S. & Dawis, R. V. The nature of objectivity with the Rasch model, Journal of Educational Measurement, 1974, 11, 163-178.

Wood, R. L., Wingersky, M. S. & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, N. J.: Educational Testing Service, 1976.

Woods, E. M. & Wiley, D. E. An application of item characteristic curve equating to single-form tests. Paper presented to Psychometric Society, Chapel Hill, N.C., 1977.

Woods, E. M. & Wiley, D. E. An application of item characteristic curve equating to item sampling packages on multi-form tests. Paper presented to American Educational Research Association, Toronto, Canada, 1978.

Wright, B. D. Sample-free test calibration and person measurement in Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116

Wright, B. D. & Stone, M. H. Best test design: A handbook for Rasch measurement. Chicago: MESA, 1979.

Table 1
Cross-Validation Statistic Values

Level	Test	Form	Equating Method								
			EQUI	LIN	TEQM1	ESTOBM1	TSEQ2	ESTOBS2	TSEQ3	ESTOBS3	TSEQ1
I	Vocabulary	X	1.60	4.30	1.84	1.70	1.23	1.07	0.92	0.87	2.49
		Y	0.30	3.04	0.79	0.86	0.24	0.57	1.03	0.15	1.96
	Quantitative Thinking	X	0.27	2.98	0.74	0.66	0.74	0.62	0.63	0.40	2.45
		Y	0.30	1.88	0.36	0.36	0.33	0.36	0.81	0.32	1.32
II	Vocabulary	X	1.43	0.94	2.09	1.63	1.69	1.37	0.86	1.73	1.57
		Y	4.09	2.40	3.42	2.55	2.90	2.40	1.78	3.22	3,11
	Quantitative Thinking	X	0.45	1.32	1.04	0.75	0.96	0.71	0.27	0.82	0.60
		Y	1.88	1.78	1.81	2.31	1.33	1.89	1.28	1.79	0.54