DOCUMENT RESUME

ED 205 544 TM 810 422

AUTHOR Engelhard, George, Jr.; Osberg, David W.
TITLE Constructing a Test Network with a Rasch Measurement Model.
PUB DATE Mar 81
NOTE 29p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (Philadelphia, March 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Elementary Secondary Education; *Equated Scores; Goodness of Fit; *Latent Trait Theory; *Mathematical Models; Tables (Data)
IDENTIFIERS *Rasch Model; Test Linking; *Vertical Equating

ABSTRACT

The purpose of this study is to present and illustrate the application of a general linear model for the analysis of test networks based on Rasch measurement models. Test networks can be used to vertically equate a set of tests which cover a wide range of difficulties. The criteria of coherence and consistency are proposed in order to assess the adequacy of the vertical equating within the test network. The method is illustrated using a set of standardized reading tests which are a part of the Achievement Series of the Comprehensive Assessment Program. (Author/BW)

CONSTRUCTING A TEST NETWORK WITH A RASCH MEASUREMENT MODEL

George Engelhard, Jr.
Chicago State University

and

David W. Ósberg
Northwestern University

Paper presented at the Eastern Educational Research Association meeting in Philadelphia, March, 1981.

ABSTRACT

The purpose of this study is to present and illustrate the application of a general linear model for the analysis of test networks based on Rasch measurement models. Test networks can be used to vertically equate a set of tests which cover a wide range of difficulties. The criteria of coherence and consistency are proposed in order to assess the adequacy of the vertical equating within the test network. The method is illustrated using a set of standardized reading tests which are a part of the Achievement Series of the Comprehensive Assessment Program (Scott, Foresman and Company, 1980).

CONSTRUCTING A TEST NETWORK WITH A RASCH MEASUREMENT MODEL

The equating of person measurements obtained on tests composed of different items is one of the major problems encountered in educational measurement. This problem arises whenever we calibrate a set of items that have a wide range of difficulties which go beyond a single individual's ability to provide meaningful responses. For example, as educators we may be interested in tracing an individual's development in reading comprehension over the elementary and secondary school years. It would be extremely difficult to develop a set of test items or a reading passage that would be appropriate for both first and twelfth graders. An item would either be so hard that everyone fails (except by chance) or so easy that everyone succeeds; both of these cases provide no useful information that can be used to calibrate the items.

In order to deal with this problem, comprehensive achievement test series designed to measure achievement over a wide age or grade range are typically composed of several levels and forms which are designed to be appropriate for selected age or grade groups. The separate levels and form of the achievement series must then be equated, so that the person ability estimates obtained from different sets of items are comparable and can be used to represent the location of the individual on one unidimensional trait that spans the age or grade levels over which we wish to trace growth or change. The goal in test equating is to step beyond the specific items contained in the separate test levels and forms

in order to obtain information on the latent trait from each individual we are interested in measuring. If an achievement test series is composed of items calibrated on a single unidimensional latent trait scale, then it becomes possible to obtain equivalent and comparable estimates of each individual's location on this latent trait regardless of form or level. The essence of the equating problem, then is to develop procedures for determining and testing the comparability of the ability estimates obtained from several different tests composed of different items over a specified difficulty range. If the forms are designed to measure the latent trait at similar ability levels, the procedure is generally called horizontal equating, e.g., alternate forms equating. The equating of measurement results obtained on tests of different levels of difficulty is called vertical equating. The purpose of this paper is to develop and illustrate a solution to the problems encountered in the vertical equating of an achievement test series based on the simplest latent trait model, the Rasch model.

Background

Various methods have been proposed as solutions to the problem of vertical equating. The problem was recognized as early as the 1920s when Thorndike pointed out that,

> With the development of group tests and tests for use with higher levels of intelligence, it is becoming more and more necessary to transmute a score obtained with one test into the score that is equivalent to it in some other test.
>
> (Thorndike, 1922, p. 29)

Thorndike "transmuted" scores using his probable error method of scaling (Thorndike, 1922; Trabue, 1916). Thurstone in a series of articles in the 1920's described his absolute scaling method which he proposed as a solution to the problem of vertical equating (Thurstone, 1925, 1927, 1928). (See Flanagan and Schwarz [1958] for an illustration of Thurstone's method of absolute scaling applied to the equating of intelligence test scores). Another solution for the vertical equating of scores is provided by the equipercentile method of vertical equating. More recently, latent trait measurement theory has been recommended as a source of solution to the "intractable" problem of equating (Lord, 1977; Marco, 1977; Wright, 1977; Rasch, 1960; Wright, 1967; Wright and Stone, 1979). A great deal of recent attention in the psychometric literature has been directed towards assessing the adequacy of the Rasch model for vertical equating and comparison of the various available methods for vertical equating (Wright, 1967; Rentz and Bashaw, 1977, 1975; Slinde and Linn, 1978, 1979; Kolen, 1980; Loyd and Hoover, 1980). These studies have lead to conflicting conclusions over the adequacy of the Rasch model for the vertical equating.

The conflicting conclusions over the adequacy of the Rasch model for vertical equating stem from two major sources. The first source of conflict involves the robustness of the Rasch model. If the test items do not fit the Rasch model, then a vertical equating based on these items will lead to an unsatisfactory equating. The key here is to use Rasch test development strategies in order to produce a set of these items which have the property of specific objectivity (Rasch, 1966a,1966b). If specific objectivity in Rasch's sense is obtained, then the person-free calibration of items can be achieved. This characteristic is always an hypothesis that must be specifically tested in each measuring situation. Slinde and Linn (1978, 1979)

and Loyd and Hoover (1980) were examining test items that were not designed to fit the Rasch model. Another source that may account for the conflicting conclusions is the problem of selecting criteria for determining the adequacy of each method for the vertical equating of a series of tests. There is no single objective criteria for comparing the results of equipercentile and latent trait equating. The traditional criterion that "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 either group are equal" proposed by Angoff (1971, p. 563) or the criterion proposed by Lord (1977, p. 128) that tests X and Y can be considered equated "... if and only if it is a matter of indifference to each examinee whether he is to take test X or test Y," may unfairly bias the conclusion about the adequacy of a method for vertical equating.

In this study, the tests that are to be equated were developed according to Rasch test development procedures. The items were piloted in a representative national sample and items were selected that fit the Rasch model. In this case, a single linking constant based on common items should theoretically provide equivalent person measurements regardless of test level or form.

The issue of what criteria to use to assess the adequacy of the vertical equating is one that requires further attention. For the purposes of this study, the adequacy of the Rasch model is defined in terms of the consistency and coherence of the linking constants within the test network. Wright (1977) in his discussion of of test networks conceived of these networks as being composed of building blocks based

on sets of three tests which form a triangle. He then suggests that tests of fit be made for each triangle in the test network. If each triangle in the test network fit the Rasch model, the three linking constants should sum to within a standard error or two of zero. The standard error of such a sum is about $3.5/(N_{12}K_{12} + N_{13}K_{13} + N_{23}K_{23})$ in which the N's are the calibration sample sizes and K's are the number of items in each link (Wright and Stone, 1979). This logic can be extended to cover other triangles of tests within the test network. The success of the test network is then assessed on the basis of a series of consistency checks performed on all possible triangles. In complete test networks, i.e., sets of common items are available for all tests in the network, this solution and procedure seems workable. In constructing test networks with a Rasch measurement model designed to vertically equate tests over a wide range of difficulties, situations are encountered where common items are not available for linking every form and level. This leads to incomplete test networks in which no direct information is available to estimate the needed linking constants. This is not an insurmountable problem and procedures have been developed to estimate links where no direct information is available (Wright and Stone, 1979). The method proposed in this paper is not intended to supplant these methods, but it is intended to provide a comprehensive approach with information about the overall coherence and consistency of a test network, while at the same time providing a method for estimating missing linking constants and providing a test of fit for each of the observed linking constants in the test network. The method can be used with complete and incomplete test networks.

The method proposed in this paper begins with the matrix of linking constants which have already been developed for the tests where common items are

available. The procedure follows a general method based on the general linear model for handling missing data outlined by Horst (1941) and discussed by Gulliksen (1956) and Bock and Jones (1968). This procedure has primarily been used with paired comparison data (Thurstone, 1927). Since the matrices produced in paired comparison experiments are similar in form and structure to the matrices obtained in test networks, this procedure suggests itself as a useful approach to the examination of the overall consistency and coherence of the test network, as well as a useful approach to the estimation of missing linking constants in incomplete test networks.

The purpose of this paper is to illustrate the application of this general linear model as a technique for examining the fit of a test network which can be used as an alternative criterion for assessing whether or not the Rasch model provides an adequate solution to the problem of vertical equating. The assumption is made that if the items within each test fit the Rasch model (an assumption that is explicitly tested), then a single linking constant provides sufficient information for obtaining equivalent person ability estimates regardless of test or form. The problem then is to assess the coherence and consistency of the network based on these linking constants using the general linear model proposed in this paper. If the observed linking constants fit the model, then the criterion of consistency is met and an adequate vertical equating has been accomplished.

Method

A General Linear Model for Examining Test Network

Let λ_{ij} represent the linking constant for equating tests i and j. This linking constant is a function of the difference between the difficulties of

test i, δ_i, and test j, δ_j. This can be written as,

$$\lambda_{ij} = \delta_i - \delta_j + e_{ij} \quad , \tag{1}$$

where e_{ij} represents a random error component. The entire matrix of linking constants for m forms or tests ($i = 1, \ldots, m$; $j = 1, \ldots, m$) can be expressed conveniently in matrix form as,

$$\underline{\lambda} = A\underline{\delta} + \underline{e} \quad , \tag{2}$$

where $\underline{\lambda}$ is a column vector of the $(m(m-1))/2$ observed linking constants, ordered by their subscripts (λ_{12}, λ_{13}, ..., λ_{1m}, λ_{23}, λ_{24}, ..., etc.); $\underline{\delta}$ is a vector of m test difficulties (δ_1, δ_2, ..., δ_m); A is a matrix that has the following form

$$A = \begin{bmatrix}
1 & -1 & 0 & 0 & \ldots & 0 & 0 \\
1 & 0 & -1 & 0 & \ldots & 0 & 0 \\
1 & 0 & 0 & -1 & \ldots & 0 & 0 \\
 & & & \vdots & & & \\
1 & 0 & 0 & 0 & \ldots & 0 & -1 \\
0 & 1 & -1 & 0 & \ldots & 0 & 0 \\
 & & & \vdots & & & \\
0 & 1 & 0 & 0 & \ldots & 0 & -1 \\
0 & 0 & 1 & -1 & \ldots & 0 & 0 \\
 & & & \vdots & & & \\
0 & 0 & 1 & 0 & \ldots & 0 & -1 \\
 & & & \vdots & & & \\
0 & 0 & 0 & 0 & \ldots & 1 & -1
\end{bmatrix}
\begin{matrix}
\left.\vphantom{\begin{matrix}1\\1\\1\\1\end{matrix}}\right\} \\
\left.\vphantom{\begin{matrix}1\\1\\1\end{matrix}}\right\} \\
\left.\vphantom{\begin{matrix}1\\1\end{matrix}}\right\} \\
\left.\vphantom{1}\right\}
\end{matrix}
\begin{matrix}
m - 1 \\
m - 2 \\
m - 3 \\
1
\end{matrix}$$

For example, the model for 3 tests (m=3) is given by,

$$\lambda_{12} = \delta_1 - \delta_2 + \varepsilon_{12}$$
$$\lambda_{13} = \delta_1 - \delta_3 + \varepsilon_{13}$$
$$\lambda_{23} = \delta_2 - \delta_3 + \varepsilon_{23}$$

or in matrix form,

$$\begin{bmatrix} \lambda_{12} \\ \lambda_{13} \\ \lambda_{23} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{bmatrix}$$

For situations where direct information is available for the linking constants, $\underline{\delta}$ can be estimated by minimizing the error term, $\underline{\varepsilon}$, in the usual way using least squares. By introducing a diagonal matrix, D, of weights it is possible to handle incomplete test networks. The least squares solution is obtained by solving the following equations,

$$Q = (\underline{\lambda} - A\underline{\delta})' D (\underline{\lambda} - A\underline{\delta}) \qquad (3)$$

or

$$Q = \underline{\varepsilon}' D \underline{\varepsilon} ,$$

where D is a diagonal matrix with 1s for the observed links and 0s for the missing links. In the case of a complete test network, D is an identity matrix. The normal equations are given by

$$A'DA\hat{\underline{\delta}} = A'D\underline{\lambda} \qquad (4)$$

and solving for the test difficulties, $\hat{\underline{\delta}}$, we have

$$\hat{\underline{\delta}} = M^{-1} \underline{z} \quad , \tag{5}$$

where $M = A'DA$ and $\underline{z} = A'D\underline{\lambda}$. Since the matrix M is not of full rank, the easiest solution in this case is to delete the last row and column of M and to delete the last row of $\underline{z}$ and solving the following equation,

$$\hat{\underline{\delta}}^* = M^{*-1} \underline{z}^* \quad . \tag{6}$$

The values of $\hat{\underline{\delta}}^*$ are the estimated test difficulties in relation to the last test. In order to obtain estimates of the linking constants, $\hat{\underline{\lambda}}$, the following equation can be used,

$$\hat{\underline{\lambda}} = A\hat{\underline{\delta}} \quad , \tag{7}$$

where $\hat{\underline{\delta}}$ differs from $\hat{\underline{\delta}}^*$ by the adjoining of a zero to the last row of $\hat{\underline{\delta}}^*$ to represent the test difficulty of the last test which is used to anchor the test network and is zero by definition.

An observed residual, $\hat{\underline{\epsilon}}$, can then be defined as

$$\hat{\underline{\epsilon}} = \underline{\lambda} - \hat{\underline{\lambda}} \quad , \tag{8}$$

and a standardized residual defined as

$$\underline{Z} = (\hat{\underline{\epsilon}} - \underline{\epsilon}_{\cdot}) / S_{\epsilon} \quad , \tag{9}$$

where $\underline{\epsilon}_{\cdot}$ is the mean vector of the residuals and S_{ϵ} is the standard deviation of the residuals. If the data fits the model, $\underline{Z}$ is approximately normally distributed with a mean of zero and a standard deviation of one.

In order to test the fit of the data to the model, the standardized

residuals can be examined and any values greater than 2 standard errors ($z > 2$) should be examined in depth. These links are suspect and the researcher may wish to eliminate these links and re-estimate the predicted links. Another approach to the analysis of the standardized residuals is to use a rankit plot (Tukey, 1962). Basically, this involves ordering the standardized residuals from the smallest to the largest, where i is an index of these ranks and s is the number of residuals. The rankits, Ri, are equal to the standard normal deviates which correspond to the following proportions for each i,

$$(3i-1)/(3s+1) \qquad (10)$$

A rankit plot can then constructed with the standardized residuals on the vertical axis and the rankits on the horizontal axis. If the model fit the data, then this plot should be a straight 45° line. If the residual analysis indicates an acceptable fit, then an adequate vertical equating has been accomplished.

The steps in examining a test network are as follows:

1. Construct an M x M matrix with observed linking constants in the matrix.
2. Construct a $[M(M-1)]/2 \times 1$ column vector composed of all of the entries above the diagonal (which is zero) in cell subscript order.
3. Obtain solutions to equations (6), (7), (8) and (9).
4. Examine the standardized residuals and determine how well the data fits the model.
5. If the fit of the model is <u>not</u> acceptable, then eliminate misfitting links, and repeat steps 3 and 4.

6. If an acceptable fit of the model is obtained, then use the estimated linking constants obtained through equation (7) for linking the tests. This vecter provides all the linking constants, and any test can be chosen at this point as an anchor test.

Description of Test Network Development and Sample

In order to construct the test network analyzed in this paper, nine linking tests with 12 to 36 common items were developed. Each one of these linking tests contained items from at least two and as many as four forms from the Achievement Series of the Comprehensive Assessment Program (Scott, Foresman and Company, 1980), designed to measure reading achievement from pre-kindergarten through high school. The overall network is shown in Figure 1. The squares represent levels 4 through 14 in the Achievement Series; the circles represent the nine linking tests specifically created for this study; the connecting lines represent sets of common items. The appropriate levels of the nine linking tests (2,4,7,8,11,12, 15,16,19) were administered to the elementary and secondary school students in Huron County, Ohio. The total number of students tested was 3,982.

BICAL (Wright, Mead and Bell, 1979) was used to test the fit of the items within each of the 20 tests (11 from the Achievement Series plus the 9 specially created linking tests). The items fit very well, and it was not necessary to eliminate any of the items at this stage. The next step was to obtain the average difficulty differences for common items in adjacent and non-adjacent test levels which will be used as the linking constants. Plots were constructed for all of the linking items, and some of these items were eliminated (see Engelhard [1980] for a description of

this procedure). In general, the elimination of common items did not have very much of an effect on the value of the linking constant. The observed linking constants obtained for the test network are given in Table 1. Table 2 gives the number of items, number of individuals, and the standard error for each link computed by equation (11).

In order to apply the linear model to this data, a computer program was written using the matrix procedures in SAS. The observed links and missing links were listed as a 190 x 1 (20 x 19/2 = 190) column vector with zeroes for all of the missing linking constants. An A matrix was constructed and a solution obtained following the earlier outlined procedure. Table 2 gives the cell subscripts, observed links, number of items, number of individuals and standard error for each link. The standard error was obtained by the following formula:

$$SE(\lambda) \doteq 3.5/(nk)^{\frac{1}{2}} \qquad (11)$$

Table 3 gives the observed and predicted links, along with an analysis of the standardized residuals. Figure 2 gives the rankit plot of the standardized residuals.

Results

Table 3 gives the residual analysis for the test network. Figure 1 gives the rankit plot for this data. In general, the data seems to fit the model relatively well, adding support to the contention that an adequate vertical equating has been accomplished. For example, using Wright's triangular analysis of a set of three tests, we see that using the observed links for linking tests 15, 16, and 19 the value of the sum is .139 (.483 + .482 − 1.104), while using the predicted links the value of the sum is − 0.001 (.523 + .540 − 1.064). By this criterion proposed

by Wright, the estimated linking constants which are based on a consideration of all the data in the linking matrix are even more coherent. Two of the standardized residuals are greater than 2; these are the links for tests 4 and 5 (1.804) and tests 4 and 7 (.863). A re-examination of the linking plots for these tests showed a considerable amount of spread in these values which should be a straight line, so that these links were not as well defined as the other links. Since by chance with 31 observed links we would expect at the .05 level 1.5 (.05 x 31) links to be greater than 2, this result is not too unlikely, and the decision was made to keep these links and no re-estimation was calculated.

Table 4 gives estimates of the test difficulties $\hat{\delta}_i$ centered on the last test ($\hat{\delta}_{20} = 0$ by definition). In order to illustrate an alternate centering of the test network, column 3 in Table 4 gives the predicted linking constants when the test network is centered on test 6 ($\hat{\delta}_6 = 0.0$). These values can be obtained in two ways. They can be obtained from the vector of predicted linking constants ($\hat{\underline{\lambda}}$) or they can more simply be obtained by subtracting the values in $\hat{\underline{\delta}}$ from the value of $\hat{\delta}_6$. This will re-center the test network on test 6. Table 4 also gives the set of initial linking constants that were used in the preliminary calibration of the reading tests in the Achievement Series. These initial values were obtained by averaging different possible linking paths and in some cases through simply taking the shortest path between two tests and summing the necessary observed linking constants. This earlier procedure did not take into account all of the linking information that was available and the results differ from the results of the study by an average of .6 of a logit which is a significant difference.

Discussion

In this paper I have proposed a method for assessing the consistency and coherence of a test network thant can be used to vertically equate a set of tests. This paper differs from previous research on vertical equating with the Rasch model in several aspects. The first difference is that the tests that are vertically equated within the test network were developed using the Rasch model. Recent research by Slinde and Linn (1978, 1979) and Loyd and Hoover (1980) have been basically research on the robustness of the Rasch model. They have examined how well the Rasch model fits item or test data that is already available. Any equating and particularly any vertical equating that is based on misfitting items will not be completely adequate. The key in any type of equating based on the Rasch model is to have items and tests that fit the Rasch model, and therefore have the desiralble properties that are associated with specific objectivity. The fit of the items within each test, the fit of items in each link and finally the fit of the linking constant within the test network must be examined. In those cases where the data fits the model, a single linking constant provides sufficient information for the equating of person measurements obtained form tests that vary in difficulty.

The method presented in this paper is similar to the one used by Rentz and Bashaw (1975, 1977). Their matrix of linking constants is slightly different in form from the one analyzed in this paper. They have two linking constants for each test, based on separate administrations of the tests given in different time order. (It would be interesting to extend the model given in this paper to include a test of the significance of this time order effect.)

The computation of row means and the use of these means as average test difficulties that can be used as linking constants is equivalent to the values obtained by the method proposed in this paper. This paper adds to their approach by making explicit the general linear model implied by the method used by Rentz, Bashaw and Wright. By making this model explicit, it is possible to obtain tests of the fit of each linking constant within the overall test network, and also to obtain least squares estimates for the unavailable links in incomplete test networks.

This paper presents a method that extends the criterion of consistency proposed by Wright (1977) for examining these networks. Previous research on the adequacy of the Rasch model for vertical equating has either compared results from different equating methods, or divided the people into different ability groups and compared the results of the separate calibrations in each group. The problem with both of these approaches is that there is no single objective criterion or any "best" method of equating that can be used as a standard criterion for comparing equating methods. I have suggested that the criterion for assessing the adequacy of a vertical equating using the Rasch model be based on a consideration of the following conditions. The first condition for an acceptable equating is that the items within each test fit the Rasch model. (See Wright and Stone [1979] for tests of item fit.) If this is true, then a single linking constant based on common items can be used to vertically equate the tests. The second condition is that the common items used to compute the linking constant must be linearly related. A plot of the difficulties for these common items obtained from the separate tests to be equated can then be represented by a straight line with a slope of one. (See Engelhard [1980] for an example of this analysis of the fit of items to the link.)

The last condition is that the criteria of coherence and consistency of the linking constants within the test network must be met for an acceptable equating. If these three conditions are met, then an acceptable equating of the tests in the network has been realized. In the present example, the three conditions for acceptable equating were met.

References

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

Bock, R.D. and Jones, L.V. The measurement and prediction of judgement and choice. San Francisco, California: Holden-Day, 1968.

Engelhard, G. An introduction to Rasch measurement and its application to test equating in the Comprehensive Assessment Program. Paper presented at the Northern Illinois Association meeting in Bloomingdale, Illinois, May, 1980.

Flanagan, J.C. and Schwarz, P.A. Development of procedures for converting intelligence scores to a common scale. Pittsburgh: American Institute for Research, July, 1958.

Gulliksen, H. A least squares solution for paired comparisons with incomplete data. Psychometrika, 1956, 21, 125-134.

Horst, P. The prediction of personal adjustment. Social Science Research Council, No. 48, 1941.

Kolen, M.J. Comparison of traditional and latent trait theory methods for equating tests. Paper presented at the annual meeting of American Educational Research Association in Boston, April, 1980.

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Loyd, B.H. and Hoover, H.D. Vertical equating using the Rasch model. Journal of Educational Measurement, 1980, 17, 179-193.

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

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Dansmarks Paedagogiske Institute, 1960.

Rasch, G. An individualistic approach to item analysis. In Readings in mathematical social science. Edited by Lazarsfeld, P.F. and Henry, N.W. Chicago: Science Research Associates, Inc., 1966a, 89-107.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966b, 19, Part 1, 49-57.

Rentz, R.R. and Bashaw, W.L. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-179.

Rentz, R.R. and Bashaw, W.L. Equating reading tests with the Rasch model, Vol. I Final Report, Vol. II Technical Reference Tables. Athens, GA.: University of Georgia, Educational Research Laboratory, 1975. (ERIC Document Reproduction Nos. ED 127 330 through ED 127 331.

Scott, Foresman, and Company. Comprehensive Assessment Program, 1980.

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

Slinde, J.A. and 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.

Thorndike, E.L. On finding equivalent scores in tests of intelligence. Journal of Applied Psychology, 1922, 6, 29-33.

Thurstone, L.L. A method of scaling psychological and educational tests. Journal of Educational Psychology, 1925, 15, 433-451.

Thurstone, L.L. The unit of measurement in educational scales. Journal of Educational Psychology, 1927a, 505-524.

Thurstone, L.L. The method of paired comparisons for social values. Journal of Abnormal and Social Psychology, 1927, 21, 384-400.

Thurstone, L.L. Scale construction with weighted observations. Journal of Educational Psychology, 1928, 19, 441-453.

Trabue, M.R. Completion-test language scales. Contributions to Education, No. 77. New York: Columbia University, Teachers College, 1916.

Tukey, J.W. The future of data analysis. Annals of Mathematical Statistics, 1962, 33, 1-67.

Wright, B.D. Sample-free test calibration and person measurement. In the 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. and Stone, M.H. Best Test Design. Chicago: MESA Press, 1979.

Table 1. Matrix of observed linking constants.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	1.013	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2		0	.015	1.608	.773	-	2.509	-	-	-	-	-	-	-	-	-	-	-	-	-
3			0	1.301	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4				0	1.084	2.441	.863	-	-	-	-	-	-	-	-	-	-	-	-	-
5					0	-	1.684	-	-	-	-	-	-	-	-	-	-	-	-	-
6						0	.567	-	-	-	-	-	-	-	-	-	-	-	-	-
7							0	.980	-	-	2.184	-	-	-	-	-	-	-	-	-
8								0	.246	1.51	1.323	-	-	-	-	-	-	-	-	-
9									0	-	-	-	-	-	-	-	-	-	-	-
10	[Same as values above diagonal, except the values are negative]									0	.172	-	-	-	-	-	-	-	-	-
11											0	.844	-	-	1.601	-	-	-	-	-
12												0	.003	.614	.844	-	-	-	-	-
13													0	-	-	-	-	-	-	-
14														0	.510	-	-	-	-	-
15															0	.483	-	-	1.104	-
16																0	.233	.558	.482	-
17																	0	-	.287	-
18																		0	.042	-
19																			0	.865
20																				0

Table 2. Observed links, number of items, number of individuals and standard errors of observed links

Index	Cell Subscripts	Observed Links	Number of Items	Number of Individuals	Standard error
1	1,2	1.013	8	322	.069
2	2,3	.015	12	322	.056
3	2,4	1.608	19	285	.048
4	2,5	.773	12	322	.056
5	2,7	2.509	8	322	.069
6	3,4	1.310	10	285	.065
7	4,5	1.084	12	285	.060
8	4,6	2.441	11	285	.062
9	4,7	.863	20	285	.046
10	5,7	1.684	9	333	.064
11	6,7	.567	8	333	.068
12	7,8	.980	23	312	.040
13	7,11	2.184	11	328	.058
14	8,9	.246	11	312	.060
15	8,10	1.510	12	312	.057
16	8,11	1.323	24	312	.040
17	10,11	.172	12	328	.056
18	11,12	.844	24	294	.042
19	11,15	1.601	12	307	.058
20	12,13	.003	8	294	.072
21	12,14	.614	11	294	.061
22	12,15	.844	24	294	.042
23	14,15	.510	12	307	.058
24	15,16	.483	23	307	.042
25	15,19	1.104	11	307	.060
26	16,17	.233	12	597	.041
27	16,18	.558	12	597	.041
28	16,19	.482	36	597	.024
29	17,19	.287	10	1,204	.032
30	18,19	.042	12	1,204	.029
31	19,20	.865	12	1,204	.029

Table 3. Residual analysis of linking constants.

Index	Cell Subscripts	Observed Links	Predicted Links	Residual	Residual (Std)	Rankit
1	1,2	1.013	1.013	.000	-.230	-.202
2	2,3	.015	-.128	.143	.227	.643
3	2,4	1.608	1.040	.568	1.588	1.175
4	2,5	.773	1.301	-.528	-1.920	-1.645
5	2,7	2.509	2.692	-.163	-.814	1.341
6	3,4	1.310	1.167	.143	.227	.643
7	4,5	1.084	.262	.822	2.400	2.054
8	4,6	2.441	1.763	.678	1.939	1.476
9	4,7	.863	1.652	-.789	-2.754	-2.054
10	5,7	1.684	1.390	.294	.710	1.036
11	6,7	.567	-.111	.678	1.939	1.476
12	7,8	.980	.890	.090	.057	.332
13	7,11	2.184	2.274	-.090	-.516	-1.175
14	8,9	.246	.246	.000	-.230	-.202
15	8,10	1.510	1.360	.150	.248	.842
16	8,11	1.323	1.383	-.060	-.422	-.915
17	10,11	.172	.022	.150	.248	.842
18	11,12	.844	.776	.066	-.013	.253
19	11,15	1.601	1.669	-.068	-.446	-1.036
20	12,13	.003	.003	.000	-.230	-.202
21	12,14	.614	.498	.116	.141	.468
22	12,15	.844	.892	-.048	-.384	-.706
23	14,15	.510	.394	.116	.141	.468
24	15,16	.483	.523	-.040	-.358	-.583
25	15,19	1.104	1.064	.040	-.101	.151
26	16,17	.233	.245	-.012	-.267	-.468
27	16,18	.558	.528	.030	-.134	.050
28	16,19	.482	.540	-.058	-.416	-.806
29	17,19	.287	.296	-.012	-.267	-.468
30	18,19	.042	.012	.030	-.134	.050
31	19,20	.865	.865	.000	-.230	-.202
Mean		.927	.856	.072	.000	.001
Standard Deviation		.682	.703	.313	1.000	.966

Table 4. Test difficulties, re-centered links, and comparison with preliminary results of an earlier linking study.

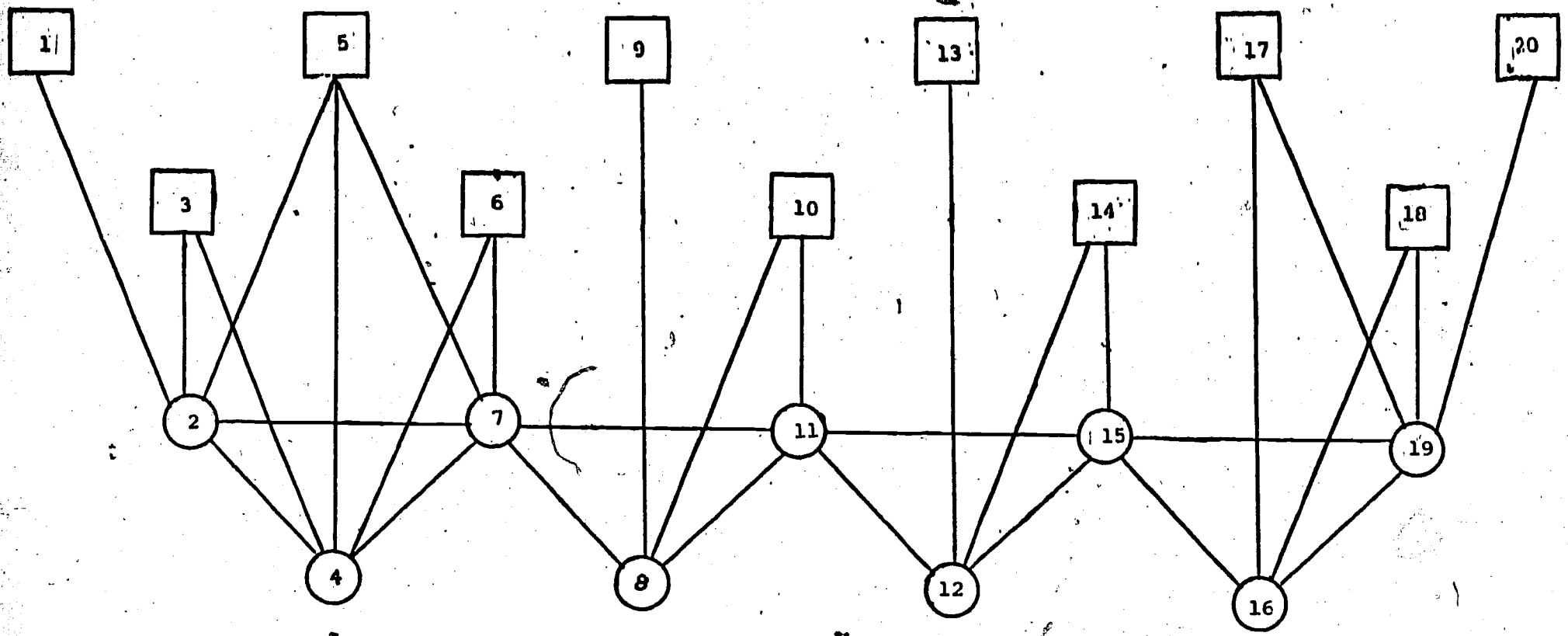
Test Difficulties (delta)	Cell Subscripts	Re-centered Links	Test[a] ID	Preliminary[b] Links	Difference
9.575	6,1	-3.815	4-7	-4.272	.457
8.562	6,2	-2.802			
8.690	6,3	-2.930	5-7	-3.625	.695
7.522	6,4	-1.763			
7.261	6,5	-1.510	6-7	-2.344	.834
5.760	6,6	0.000	7-7	0.000	.000
5.871	6,7	-.111			
4.980	6,8	.779			
4.734	6,9	1.026	7-8	.616	.410
3.620	6,10	2.140	7-9	1.696	.444
3.597	6,11	2.162			
2.821	6,12	2.939			
2.818	6,13	2.942	7-10	2.235	.707
2.323	6,14	3.437	7-11	2.635	.802
1.929	6,15	3.831			
1.405	6,16	4.354			
1.161	6,17	4.599	7-12	3.375	1.224
.877	6,18	4.883	7-13	4.177	.706
.865	6,19	4.894			
.000	6,20	5.760	7-14	5.094	.696

Note Preliminary links were only obtained for the Achievement Series and not for the specially created linking tests.

a. The test numbers used through out this paper were assigned in order to simplify the discussion in this paper. The numbers in this column show the actual form and level numbers used by Scott, Foresman and Company for their Achievement Series.

b. The values given in this column were based on a preliminary analysis of the Huron County Data used in this paper. The values currently used by Scott, Foresman and Company are based on a national representative sample. This data is not currently available.

Figure 1. Network of Reading Tests



Grade

Pre-K K 1 2 3 4 5 6 7-8 9-10 11-12

Key - Squares represent levels 4 through 14 in the Achievement Series; circles represent linking tests constructed for this study; single lines represent sets of common items

Figure 2. Rankit plot of standardized residuals.