Constructing a Test Network with a Rasch Measurement Model

George Engelhard, Jr.
University of Chicago and Chicago State University

David W. Osberg
The Riverside Publishing Company

The purpose of this study is to present and to 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 that cover a wide range of difficulties. The criteria of consistency and coherence 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 Comprehensive Assessment Program's (1981) Achievement Series.

The equating of person measurements obtained on tests composed of different items is one of the major problems encountered in psychological and educational measurement. This problem arises whenever a set of items must be equated that have a wide range of difficulties which go beyond a single individual's ability to provide meaningful responses. For example, educators may be interested in tracing an individual's growth and development in reading comprehension over the elementary and secondary school years. It would be extremely difficult, however, to develop a single test composed of reading items that would be appropriate for both first and twelfth graders. One way to deal with this problem is to create a comprehensive series of tests designed to measure achievement over a wide age or grade range. In such a series each separate test is designed to be appropriate for a certain range of ability on the latent trait continuum.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 3, Summer 1983, pp. 283–294

0146-6216/83/030283-12$1.85

The goal in test equating is to extrapolate beyond the specific items contained in the separate tests in order to obtain information on the latent trait for each individual being measured. 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 the test administered. Equating of measurements on tests designed to represent the latent trait at similar ability levels is generally called horizontal, or alternate forms, equating. Equating of measurements obtained on tests of different levels of difficulty is called vertical equating. This paper develops and illustrates a solution to some of the problems that are encountered in the vertical equating of a comprehensive 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 (1922) pointed out that

> with the development of group tests for use with higher levels of intelligence, it is becom-

> ing more and more necessary to transmute a score obtained with one test into the score that is equivalent to it in some other test. (p. 29)

Thorndike "transmuted" scores using his probable error method of scaling. Thurstone (1925, 1927, 1928) proposed that his method of absolute scaling was a solution to the problem of vertical equating. Angoff (1971) described several methods for equating psychological and educational tests. More recently, latent trait measurement theory has been recommended as a source of solutions to the "intractable" problem of equating (Haebara, 1980; Lord, 1977; Marco, 1977; Petersen, Cook, & Stocking, 1981; Rasch, 1980/1960; Wright, 1968; Wright & Stone, 1979).

The Rasch model is the simplest of the latent trait measurement models (Wright & Stone, 1979). If the test data fit the Rasch model, the vertical equating of tests can be accomplished with a single linking constant based on common items within the tests to be equated. A number of studies have examined the use of the Rasch model for vertical equating (Guskey, 1981; Loyd & Hoover, 1980; Rentz & Bashaw, 1977; Slinde & Linn, 1978, 1979; Wright, 1968). These studies have lead to conflicting conclusions over the adequacy of the Rasch model for vertical equating.

The issue of what criteria to use to assess the adequacy of a vertical equating is one that requires further attention. It is beyond the scope of this study to address the issue by comparing results based on different methods and different latent trait models. Rather, for the purpose of this study, the adequacy of a vertical equating based on a Rasch measurement model will be defined in terms of the consistency and coherence of the linking constants within a test network.

Wright (1977) described the development of test networks using the Rasch model and outlined a method for examining them. Basically, he suggested that a series of consistency checks be performed using the linking constants for each set of three tests within the test network. The success of the test network and the vertical equating would then be judged by analyzing the magnitudes and directions of these triangle sums (Wright, 1977). This method is illustrated in Wright and Stone (1979).

The model proposed in this study begins with the matrix of linking constants that have already been developed using the Rasch model. The construction of these observed linking constants has been described in Wright and Stone (1979) and also in Guskey (1981). The procedure is based on a general linear model for handling missing data outlined by Horst (1941). Gulliksen (1956) and Bock and Jones (1968) have applied a similar model to the analysis of paired comparison data. 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 a test network.

Purpose

The purpose of this study is to illustrate the application of this linear model as a procedure for examining the fit of linking constants within a test network, which can be used as an additional criterion for assessing whether or not the Rasch model provides an adequate solution to the problem of vertical equating. The assumption is made that when the items within each test fit the Rasch model, a single linking constant provides sufficient information for obtaining equivalent person ability estimates regardless of test. The problem then is to assess the consistency and coherence of the network based on these linking constants using the general linear model proposed in this study. If the observed linking constants fit the model, then the criterion of consistency is met and additional support is provided for the adequacy of the vertical equating using a Rasch measurement model.

Method

A General Linear Model for Examining a 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

$$\lambda_{ij} = \delta_i - \delta_j + \varepsilon_{ij} , \qquad [1]$$

where ε_{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 follows:

$$\boldsymbol{\lambda} = \mathbf{A}\boldsymbol{\delta} + \boldsymbol{\epsilon} , \qquad [2]$$

where $\boldsymbol{\lambda}$ is a column vector of $m(m-1)/2$ observed linking constants, ordered by their subscripts (λ_{12}, $\lambda_{13}, \ldots, \lambda'_{1m}, \lambda_{23}, \lambda_{24}, \ldots$, and so forth; $\boldsymbol{\delta}$ is a vector of m test difficulties ($\delta_1, \delta_2, \ldots, \delta_m$); $\mathbf{A}$ is a $m(m-1)/2$ by m matrix that has the following form:

$$\mathbf{A} = \begin{array}{c} \left[\begin{array}{cccccc} 1 & -1 & 0 & \ldots & 0 & 0 \\ 1 & 0 & -1 & \ldots & 0 & 0 \\ & & \vdots & & & \\ 1 & 0 & 0 & \ldots & 0 & -1 \\ 0 & 1 & -1 & \ldots & 0 & 0 \\ & & \vdots & & & \\ 0 & 1 & 0 & \ldots & 0 & -1 \\ & & \vdots & & & \\ 0 & 0 & 0 & \ldots & 1 & -1 \end{array}\right] \end{array} \begin{array}{l} \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} m-1 \\ \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} m-2 \\ \\ \left.\right\} 1 \end{array} . \qquad [3]$$

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

$$\begin{aligned} \lambda_{12} &= \delta_1 - \delta_2 + \varepsilon_{12} \\ \lambda_{13} &= \delta_1 - \delta_3 + \varepsilon_{13} \\ \lambda_{23} &= \delta_2 - \delta_3 + \varepsilon_{23} , \end{aligned} \qquad [4]$$

or in matrix form,

$$\begin{bmatrix} \lambda_{12} \\ \lambda_{13} \\ \lambda_{23} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{bmatrix} . \qquad [5]$$

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

$$\mathbf{Q} = (\boldsymbol{\lambda} - \mathbf{A}\boldsymbol{\delta})' \, \mathbf{D}(\boldsymbol{\lambda} - \mathbf{A}\boldsymbol{\delta}) \qquad [6]$$

or

$$\mathbf{Q} = \boldsymbol{\epsilon}'\mathbf{D}\boldsymbol{\epsilon} , \qquad [7]$$

where $\mathbf{D}$ is a diagonal matrix with ones for the observed links and zeros for the missing links. In the case of a complete test network, $\mathbf{D}$ is an identity matrix. The normal equations are given by

$$\mathbf{A}'\mathbf{D}\mathbf{A}\hat{\boldsymbol{\delta}} = \mathbf{A}'\mathbf{D}\boldsymbol{\lambda} , \qquad [8]$$

and solving for the test difficulties, $\hat{\boldsymbol{\delta}}$, gives

$$\hat{\boldsymbol{\delta}} = \mathbf{M}^{-1}\,\mathbf{Z} , \qquad [9]$$

where $\mathbf{M} = \mathbf{A}'\mathbf{D}\mathbf{A}$ and $\mathbf{Z} = \mathbf{A}'\mathbf{D}\boldsymbol{\lambda}$. Since the matrix $\mathbf{M}$ is not of full rank, the easiest solution in this case is to delete the last row and column of $\mathbf{M}$ and to delete the last row of $\mathbf{Z}$ and then solve the following equation,

$$\hat{\boldsymbol{\delta}}^* = \mathbf{M}^{*-1}\,\mathbf{Z}^* . \qquad [10]$$

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

$$\hat{\boldsymbol{\lambda}} = \mathbf{A}\hat{\boldsymbol{\delta}} , \qquad [11]$$

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

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

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}} , \qquad [12]$$

and a standardized residual defined as

$$\mathbf{E} = (\hat{\boldsymbol{\epsilon}} - \hat{\varepsilon}.)/S_\varepsilon , \qquad [13]$$

where $\hat{\varepsilon}.$ is the mean of the vector of residuals and S_ε is the standard deviation of the residuals. If the data fit the model, the values in the vector $\mathbf{E}$ should be 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 two standard errors examined in depth. These misfitting links may be eliminated and the predicted links reestimated. Another approach to the analysis of standardized residuals involves the use of 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, R_i, are equal to the standard normal deviates which correspond to the following proportions for each i,

$$\mathbf{R}_i = (3i - 1)/(3s+1) \, . \qquad [14]$$

A rankit plot can then be constructed with the standardized residuals on the vertical axis and the rankits on the horizontal axis. If the data fit the model, then this plot should be a straight line with a 45-degree angle. If the residual analysis indicates an acceptable fit, then additional evidence is provided for the adequacy of the vertical equating with the test network.

The steps in examining a test network can be summarized as follows:

1. Construct an $m \times m$ matrix of observed linking constants.
2. Construct a $[m(m-1)/2] \times 1$ column vector composed of all the entries above the diagonal in cell subscript order.
3. Solve Equations 10, 11, 12, and 13.
4. Examine the standardized residuals and determine how well the data fit the model.
5. If the fit of the data is *not* acceptable, then eliminate misfitting links, and repeat Steps 3 and 4.
6. If an acceptable fit of the data is obtained, then use the estimated linking constants obtained through Equation 11. This vector provides all the linking constants and any test can be chosen at this point as an anchor test.

Development of the Test Network

Nine linking tests with 12 to 36 common items were developed in order to construct the test network analyzed in this study. Each one of these linking tests contained items from at least two and as many as four forms from the Comprehensive Assessment Program (1981) Achievement Series. The reading tests, which are designed to assess reading achievement from prekindergarten through high school, were selected for this study.

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 that were specifically created for this study, and the connecting lines represent sets of common items. The appropriate levels of the nine linking tests (2, 4, 7, 8, 11, 15, 16, 19) were administered to the elementary and secondary school children in Huron County, Ohio. The total number of students tested was 3,982.

BICAL (Wright, Mead, & Bell, 1979) was used to test the fit of the items within each of the 20 tests (11 from the Achievement Series and the 9 specially created linking tests). The items fit the model very well, and it was not necessary to eliminate any items at this stage. The next step was to obtain the average difficulties of the differences for common items in adjacent and nonadjacent tests which were used as linking constants. Plots were constructed for all the common linking items, and some items were eliminated. In general, the elimination of misfitting items did not have a very great effect on the value of the linking constants. The observed linking constants for the test network are given in Table 1.

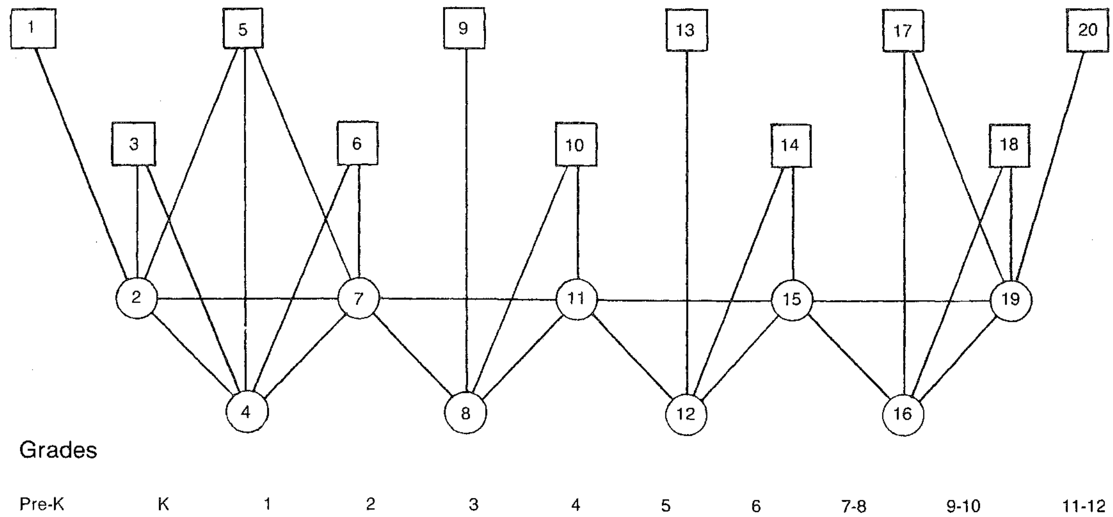
In order to apply the general linear model proposed in this study, a computer program was written using the matrix procedures in SAS. The observed linking constants and the missing linking constants were loaded into a 190×1 column vector with zeros indicating the missing linking constants. An $\mathbf{A}$ matrix was constructed and a solution obtained following the procedures outlined earlier. Table 2 gives the cell subscripts, observed linking constants, number of items, number of individuals, and the standard error for each linking constant. The standard error was computed by the following equation given by Wright and Stone (1979, p. 96):

$$\mathrm{SE}(\boldsymbol{\lambda}) \simeq 3.5/(nk)^{1/2} \, . \qquad [15]$$

Results

In general, the data seem to fit the model relatively well, adding support to the contention that an adequate vertical equating has been accomplished. Table 3 gives the observed and predicted linking constants along with an analysis of the standardized residuals. Figure 2 gives the rankit plot

Figure 1
Network of Reading Achievement Tests



of the standardized residuals. Using the method proposed by Wright (1977), the observed linking constants for the three tests—15, 16, and 19—sum to .139 (.483 + .482 − 1.104); whereas using the predicted linking constants, the sum is −.001 (.523 + .540 − 1.064). By the criterion proposed by Wright (1977), the estimated linking constants that are based on a consideration of all the data in the linking network are even more consistent. Two of the standardized residuals are greater than 2; these are the linking constants for Tests 4 and 5 (1.804) and for Tests 4 and 7 (.863). A reexamination of the plots of linking constants for these tests showed a considerable amount of spread in these values, which should be linearly related. Therefore, these linking constants did not seem to be as well defined as the other links. A decision was made, however, to keep these linking constants in the model because there is a reasonably high probability of finding two misfitting links by chance even when the model is appropriate.

Table 4 gives the estimates of the test difficulties centered on the last test, which is zero 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. These values can be obtained in two ways. They can be obtained from the vector of predicted linking constants $\hat{\boldsymbol{\lambda}}$ or, more simply, by subtracting the values that are needed to recenter the test network.

Table 4 also gives the set of initial linking constants that were used in the preliminary calibration of the reading tests of the Achievement Series. These initial values were obtained by averaging different possible linking constants based on different possible paths between tests. In some cases the linking constants were obtained simply by 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 current study by an average of .6 logits.

Discussion

The purpose of this study was to introduce a general linear model that can be used to examine the consistency and coherence of a test network based on a Rasch measurement model. This linear model extends the criterion of consistency proposed by Wright (1977) for examining test networks. The model provides a comprehensive ap-

Table 1
Matrix of Observed Linking Constants
(values below the diagonal are the same except for sign)

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	-	1.013	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2		-	.015	1.608	.773	-	2.509	-	-	-	-	-	-	-	-	-	-	-	-	-
3			-	1.301	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4				-	1.084	2.44	.863	-	-	-	-	-	-	-	-	-	-	-	-	-
5					-	-	1.684	-	-	-	-	-	-	-	-	-	-	-	-	-
6						-	.567	-	-	-	-	-	-	-	-	-	-	-	-	-
7							-	.980	-	-	2.184	-	-	-	-	-	-	-	-	-
8								-	.246	1.51	1.323	-	-	-	-	-	-	-	-	-
9									-	-	-	-	-	-	-	-	-	-	-	-
10										-	.172	-	-	-	-	-	-	-	-	-
11											-	.844	-	-	1.601	-	-	-	-	-
12												-	.003	.614	.844	-	-	-	-	-
13													-	-	-	-	-	-	-	-
14														-	.510	-	-	-	-	-
15															-	.483	-	-	1.104	-
16																-	.233	.558	.482	-
17																	-	-	.287	-
18																		-	.042	-
19																			-	.865
20																				-

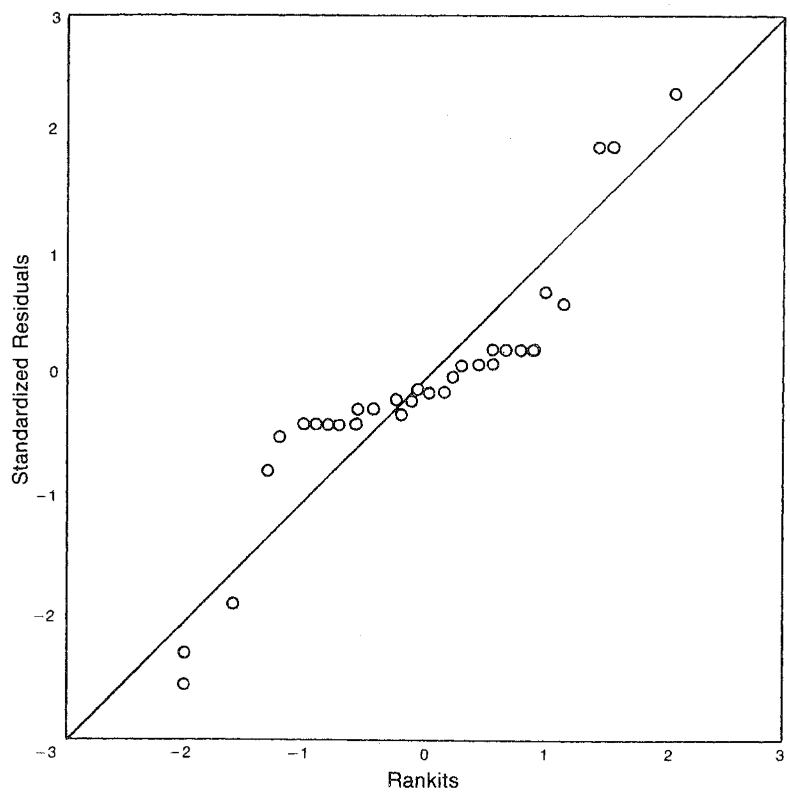
Table 2
Description of Observed Linking Constants

Cell Subscripts	Linking Constants	Number of Items	Number of Individuals	Standard Errors
1,2	1.013	8	322	.069
2,3	.015	12	322	.056
2,4	1.608	19	285	.048
2,5	.773	12	322	.056
2,7	2.509	8	322	.069
3,4	1.310	10	285	.065
4,5	1.084	12	285	.060
4,6	2.441	11	285	.062
4,7	.863	20	285	.046
5,7	1.684	9	333	.064
6,7	.567	8	333	.068
7,8	.980	23	312	.040
7,11	2.184	11	328	.058
8,9	.246	11	312	.060
8,10	1.510	12	312	.057
8,11	1.323	24	312	.040
10,11	.172	12	328	.056
11,12	.844	24	294	.042
11,15	1.601	12	307	.058
12,13	.003	8	294	.072
12,14	.614	11	294	.061
12,15	.844	24	294	.042
14,15	.510	12	307	.058
15,16	.483	23	307	.042
15,19	1.104	11	307	.060
16,17	.233	12	597	.041
16,18	.588	12	597	.041
16,19	.482	36	597	.024
17,19	.287	10	1,204	.032
18,19	.042	12	1,204	.029
19,20	.865	12	1,204	.029

Table 3
Residual Analysis of Linking Constants

Cell Subscripts	Linking Constants Observed	Linking Constants Predicted	Standardized Residuals	Rankits
1,2	1.013	1.013	-.230	-.202
2,3	.015	-.128	.227	.643
2,4	1.608	1.040	1.588	1.175
2,5	.773	1.301	-1.920	-1.645
2,7	2.509	2.692	-.814	1.341
3,4	1.310	1.167	.227	.643
4,5	1.084	.262	2.400	2.054
4,6	2.441	1.763	1.939	1.476
4,7	.863	1.652	-2.754	-2.054
5,7	1.684	1.390	.710	1.036
6,7	.567	-.111	1.939	1.476
7,8	.980	.890	.057	.332
7,11	2.184	2.274	-.516	-1.175
8,9	.246	.246	-.230	-.202
8,10	1.510	1.360	.248	.842
8,11	1.323	1.383	-.422	-.915
10,11	.172	.022	.248	.842
11,12	.844	.776	-.013	.253
11,15	1.601	1.669	-.446	-1.036
12,13	.003	.003	-.230	-.202
12,14	.614	.498	.141	.468
12,15	.844	.892	-.384	-.706
14,15	.510	.394	.141	.468
15,16	.483	.523	-.358	-.583
15,19	1.104	1.064	-.101	.151
16,17	.233	.245	-.267	-.468
16,18	.558	.528	-.134	.050
16,19	.482	.540	-.416	-.806
17,19	.287	.296	-.267	-.468
18,19	.042	.012	-.134	.050
19,20	.865	.865	-.230	-.202
Mean	.927	.856	.000	.001
Standard deviation	.682	.703	1.000	.966

Figure 2
Rankit Plot of Standardized Residuals



proach for examining the overall consistency and coherence of a test network utilizing all of the information available in the test network. This is especially important when the vertically equated tests are designed to cover a wide range of ability. The model provides a simple test of fit for each of the observed linking constants in the test network as well as a straightforward way of estimating missing linking constants in incomplete test networks.

Previous research on the adequacy of the Rasch model for vertical equating has either compared results from different equating methods (e.g., Guskey, 1981) or has divided the sample of people into different ability groups and compared the results of separate calibrations obtained in each group (Wright & Stone, 1979). The key in any type of test equating based on the Rasch model is to have items and tests that fit the model and therefore to have the desirable properties associated with specific objectivity (Rasch, 1977). The fit of items within each test, the fit of items in each link, and finally the fit of the linking constants within the test network must be examined.

The results of this study suggest that the criteria for assessing the adequacy of a vertical equating using a Rasch measurement model must be based on a consideration of the following three conditions. The first condition for an acceptable equating is that the items within each test fit the Rasch model. The second condition is that the common items used to compute the linking constants must be linearly related. A plot of the difficulties for these common items can be used to test this condition. The last condition is that the criteria of consistency and coherence of the linking constant within the test network must be met for an acceptable vertical equating. If these three conditions are met, then additional support is provided for the contention

Table 4
Estimated Test Difficulties and Comparison with Preliminary Linking Constants

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

*Preliminary linking constants were only available for the tests in the Achievement Series.

that an adequate vertical equating has been accomplished.

The issue of how to vertically equate psychological and educational tests over a wide age or grade range is a major problem. The idea that a single unidimensional scale can be constructed using a series of vertically equated tests is relatively reasonable from a psychological perspective, but the psychometric problems encountered when vertically equating tests over a wide range should not be minimized. The stability over time of the linking constants in test networks is one issue that requires further attention. Test networks provide a practical solution to the problem, but not necessarily the final solution.

Suggestions for Future Research

There are several important areas for future research that are suggested by this study. Some of the areas are as follows:

1. If earlier results comparing the Rasch model with other methods are recomputed using more consistent linking constants based on the model

proposed in this study, will the conclusions be the same? The differences found between observed and predicted linking constants in this study suggest that the revised estimates of ability might be substantially different.

2. Is it possible to develop even more accurate linking constants by using a weighted least squares model with the standard error of the linking constants incorporated into the weight matrix?
3. Can the current model be extended to allow the testing of a time effect on the linking constants, the effect of different groups or settings, or to include cases where there are multiple linking constants available for each test in the network?
4. Is it possible to extend the current model to apply directly to items rather than tests? How would this method compare with other methods of item banking?

References

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

Bock, R. D., & Jones, L. V. *The measurement and prediction of judgment and choice*. San Francisco CA: Holden-Day, 1968.

Comprehensive Assessment Program. *Achievement Series. Technical manual, Forms A and B*. Glenview IL: Scott, Foresman, & Company, 1981.

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

Guskey, T. R. Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. *Applied Psychological Measurement*, 1981, *5*, 187–201.

Haebara, T. *Equating logistic ability scales by a weighted least squares method*. Iowa City IA: The University of Iowa, 1980. (ERIC Document Reproduction Service No. ED 193 300).

Horst, P. *The prediction of personal adjustment*. New York: Social Science Research Council, 1941. (No. 48)

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

Loyd, B. H., & 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.

Petersen, N. S., Cook, L. L., & Stocking, M. L. *IRT versus conventional equating methods: A comparative study of scale stability*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, March 1981.

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press, 1980. (Originally published, Copenhagen: Danmarks Paedagogiske Institut, 1960.)

Rasch, G. On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 1977, *14*, 58–94.

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

Slinde, J. A., & Linn, R. L. An exploration of the Rasch model for the problems 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.

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*, 1927, *18*, 505–524.

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

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 *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton NJ: Educational Testing Service, 1978.

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

Wright, B. D., Mead, R. J., & Bell, S. R. *BICAL: Calibrating items with the Rasch model* (Research Memorandum No. 23B). University of Chicago, Department of Education, Statistical Laboratory, 1979.

Wright, B. D., & Stone, M. H. *Best test design*. Chicago: MESA Press, 1979.

Acknowledgments

This research was conducted while both authors were measurement consultants for Scott, Foresman, and Company. An earlier version of this study was presented at the Eastern Educational Research Association meeting in Philadelphia, March 1981.

Author's Address

Send requests for reprints or further information to George Engelhard, Jr., Office of Institutional Research and Evaluation, Chicago State University, Ninety-Fifth Street at King Drive, Chicago IL 60628, U.S.A.