

USING A LATENT TRAIT MODEL TO VERTICALLY EQUATE EDUCATIONAL TESTS

Geofferey N. Masters
University of Melbourne
and

Leila T. Mossenson
Education Department of Western Australia

ABSTRACT

The use of the Rasch model to vertically equate a set of fourteen reading comprehension tests is described. This study differs from most previous equating studies in that the equated tests were first constructed to conform to the Rasch model, and groups of common-persons rather than sets of common-items are used for the equating. Procedures for analyzing the internal consistency of individual links, for equating the fourteen test forms, and for analyzing the coherence of the test bank as a whole are described and illustrated.

Of the procedures that have been proposed for the vertical equating of educational tests, the most promising appear to be those based on latent trait methods (Lord, 1980; Wright, 1977). To date, most applications of these methods have used Rasch's dichotomous model (Rasch, 1960/1980). However, the utility of this model for vertical equating continues to be debated (e.g., Divgi, 1981; Guskey, 1981; Gustafsson, 1979; Holmes, 1982; Loyd and Hoover, 1980; Rentz and Bashaw, 1975; Slinde and Linn, 1978, 1979a, 1979b).

This paper describes the vertical equating of a set of fourteen reading comprehension tests. It differs from previous equating studies in several important respects. First, most previous studies have attempted to equate existing standardized tests which have not been constructed to conform to the latent trait model subsequently used for the equating. The tests equated here (the *Tests of Reading Comprehension*) were constructed to approximate the measurement requirements of Rasch's dichotomous model. Items which displayed poor fit to this model were either discarded or revised during test development (Mossenson and Masters, 1983).

Second, because each of these reading tests uses a different passage of writing, no item occurs in more than one test. This means that the usual approach of embedding a set of items in both an easier and a harder test and then using these "link" items to equate the two tests is not feasible. Instead, the *Tests of Reading Comprehension* are equated by administering a pair of tests to each student in the calibration sample and then equating these two tests on the basis of the performances of those students who took both tests. In other words, the vertical equating is based on groups of "common-persons" rather than on sets of "common-items".

Third, while some previous equating studies have used tests which assess a broad range of basic skills, the *Tests of Reading Comprehension* were constructed to measure a rather narrowly defined construct. (A deliberate attempt was made not to assess word knowledge as part of reading comprehension, for example). As a consequence, the intention of unidimensionality may be better approximated in these fourteen tests than in the tests used in some previous studies.

Of particular interest in this study is the effectiveness of using groups of common-persons to link pairs of tests. Procedures for estimating the relative difficulties of a pair of test forms, for analyzing the internal consistency of individual links, for equating a set of tests, and for assessing the coherence of an entire linking structure are described and illustrated.

THE LINKING DESIGN

The *Tests of Reading Comprehension* (Mossenson and Masters, 1983) are a set of fourteen reading tests constructed for use with students in their third to tenth years of school. Each test consists of a passage of writing approximately two pages in length accompanied by a "retelling" of this passage in different words. This retelling contains gaps corresponding to details contained in the original passage. Readers who understand the passage should be able to fill in these gaps using one or more of their own words. Each insertion is then scored

either right or wrong.

The intention underlying these tests is that teachers will be able to choose tests appropriate to the reading levels of individual students and to administer different tests at different times in a student's reading development. This should be fairer than administering a single standardized reading test to all students, some of whom are likely to find it much too easy and others, much too hard. Before performances on different tests can be compared, however, these fourteen tests must be equated.

The equating of the *Tests of Reading Comprehension* was based on a sample of 2698 third to tenth grade students in Western Australia in November 1982. The linking design used to equate the tests is shown in Figure 1. The fourteen test forms are labelled A to N; the easier forms are on the left, the harder forms are on the right. Each line segment connecting a pair of tests represents a group of about 100 students who took that pair of forms. (The exceptions are the line segments joining Forms A and B and Forms M and N which each represent about 200 students). The numbers on the lines show the grades from which these students were drawn.

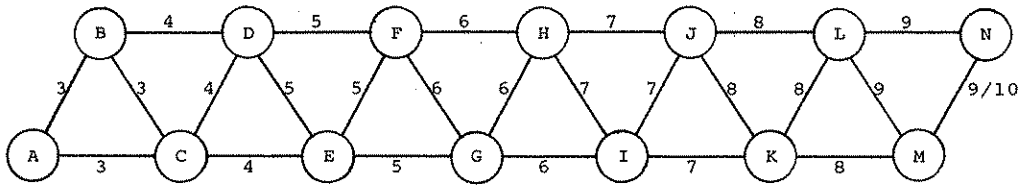


Figure 1. The linking design.

This linking design provides a number of direct and indirect estimates of the relative difficulties of these fourteen forms. The relative difficulties of Forms F and G, for example, can be estimated directly from the performances of the group of sixth grade students who took this pair of forms, or indirectly through the F-E-G and F-H-G chains. The availability of these indirect estimates provides a means of analyzing the fit of the direct estimate (i.e., the F-G link) into the linking structure as a whole.

WITHIN-LINK ANALYSIS

For each student in the calibration sample two ability estimates have been obtained--one from each of the forms taken by that student. Because each of these estimates is expressed with respect to the mean difficulty of the items in that form, a comparison of each student's two ability estimates provides an estimate of the relative difficulties of the two test forms. This is illustrated in Figure 2 where the two estimates for each of the students in the K-M link are plotted against each other. (Not included in Figure 2 are students who made a perfect or zero score on either Form K or Form M and so did not receive an ability estimate on one form, and students who failed to complete either of the two test forms).

If the two ability estimates for the n'th person in the K-M link are denoted b_{nK} and b_{nM} , then the relative difficulties of Forms K and M can be estimated by averaging the difference ($b_{nK} - b_{nM}$) over all N persons in the K-M link:

$$t'_{KM} = \frac{\sum_{n=1}^N (b_{nK} - b_{nM})}{N} \quad [1]$$

Alternatively, because some abilities are estimated more precisely than others, an improved estimate of the shift needed to equate Forms K and M may result if, in calculating this average, more weight is given to the more precisely estimated abilities¹. This can be achieved using

$$t_{KM} = \frac{\sum_{n=1}^N (b_{nK} - b_{nM}) / W_{nKM}}{\sum_{n=1}^N (1 / W_{nKM})} \quad [2]$$

¹A more complete discussion of the statistics used here is provided by Wright and Bell (1981).

where $W_{nKM} = s_{nK}^2 + s_{nM}^2$, and s_{nK} and s_{nM} are the errors of measurement associated with estimates b_{nK} and b_{nM} . The standard error of shift estimate t_{KM} is

$$s_{KM} = \left(\sum_{n=1}^N W_{nKM} / N \right)^{1/2} / N \quad [3]$$

In the case of link K-M, the estimated shift t_{KM} is .83 logits with standard error .02. In other words, Form K is estimated to be .83 logits easier than Form M. This difference must be added to the Form M abilities to bring them on to the same scale as the Form K abilities.

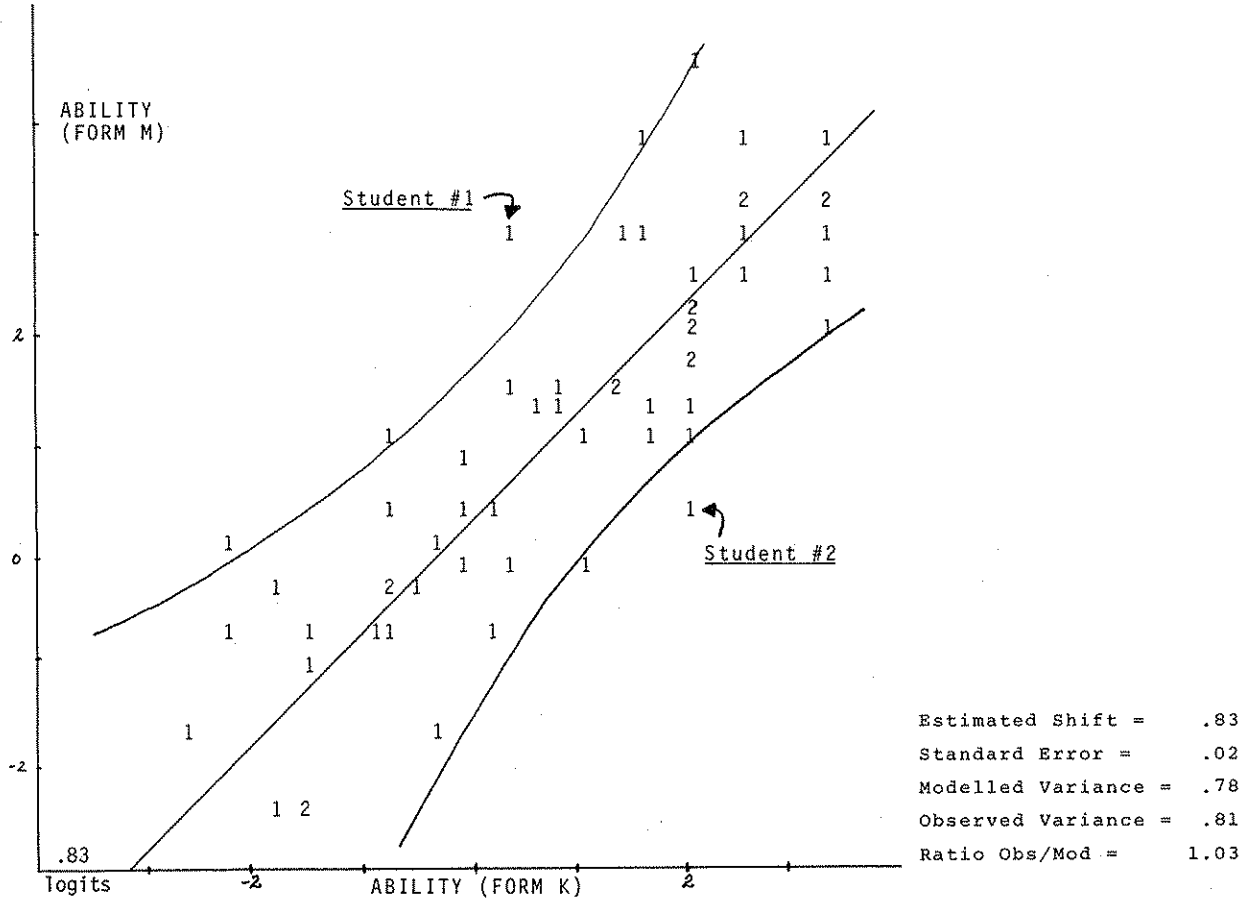


Figure 2. Link K-M.

Figure 2 also provides a graphical check on the internal consistency of link K-M. The diagonal line in Figure 2 is shifted .83 logits to the right of the identity line and a 95 percent confidence band has been added. (The procedure for constructing this confidence interval is described by Wright and Masters, 1982). Two students are some distance outside this band. Given their performances on Form K, Student #1 has performed surprisingly well on Form M, and Student #2 has performed surprisingly poorly. These two outliers provide contradictory evidence concerning the relative difficulties of Forms K and M and so lower the internal consistency of this link. The test records for these two students might be examined to see if a reason can be found for their discrepant ability estimates. If a reason can be found for a student's surprisingly low performance on one test, then rather than averaging the two available estimates for that student, it may be fairer to discount his/her questionable ability estimate altogether.

A statistical check on the internal consistency of link K-M is based on a comparison of the observed variance of the distribution of points about the diagonal line in Figure 2 with their modelled variance about this line. (In other words, on a comparison of the observed and modelled variances of the $(b_{nK} - b_{nM})$ differences after adjusting one set of estimates for the difference in scale origin). The observed variance of the $(b_{nK} - b_{nM})$ differences is calculated as

$$S_{KM}^2 = \frac{\sum_{n=1}^N (b_{nK} - b_{nM})^2 / W_{nKM}}{\sum_{n=1}^N 1 / W_{nKM}} - \frac{\sum_{n=1}^N (b_{nK} - b_{nM}) / W_{nKM}}{\sum_{n=1}^N 1 / W_{nKM}}^2 \quad [4]$$

where once again greatest weight is given to those differences which are most precisely estimated. The modelled variance of the $(b_{nK} - b_{nM})$ differences is given by

$$V_{KM} = \frac{\sum_{n=1}^N W_{nKM}}{N} \quad [5]$$

The ratio of the observed variance S^2 to modelled variance V provides an index of the internal consistency of link K-M. For this link, the ratio is $.81/.78 = 1.03$ which is close to its expected value of 1.00.

Weighted shift estimates and their estimation errors have been calculated for all twenty-five links in this design and are shown in Table 1. Also shown are the observed and modelled variances of the $(b_{nX} - b_{nY})$ differences for each link X-Y, and their ratio. An inspection of the variance ratios on the far right of Table 1 shows a greater variation in the $(b_{nX} - b_{nY})$ differences than has been modelled. On average, the observed variances are about 30 percent larger than the modelled variances.

TABLE 1
WITHIN-LINK ANALYSIS

| LINK | ESTIMATED SHIFT t | ERROR s | MODELLED VARIANCE V | OBSERVED VARIANCE S ² | RATIO |
|------|-------------------------|------------|---------------------------|--|-------|
| A-B | -.29 | .01 | .79 | 1.04 | 1.32 |
| A-C | -.29 | .02 | .73 | 1.21 | 1.65 |
| B-C | .18 | .02 | .85 | 1.08 | 1.28 |
| B-D | 1.25 | .02 | .84 | 1.09 | 1.30 |
| C-D | .83 | .02 | .79 | 1.11 | 1.40 |
| C-E | .91 | .01 | .81 | 1.11 | 1.36 |
| D-E | .28 | .01 | .80 | .99 | 1.23 |
| D-F | .61 | .01 | .67 | 1.00 | 1.50 |
| E-F | .49 | .01 | .67 | 1.00 | 1.48 |
| E-G | .08 | .01 | .85 | .98 | 1.15 |
| F-G | .39 | .02 | .96 | 1.15 | 1.19 |
| F-H | .33 | .02 | .73 | 1.12 | 1.53 |
| G-H | .39 | .01 | .83 | 1.22 | 1.47 |
| G-I | 1.42 | .01 | .77 | .80 | 1.05 |
| H-I | 1.09 | .01 | .89 | 1.13 | 1.28 |
| H-J | .10 | .02 | .81 | 1.31 | 1.62 |
| I-J | -.89 | .01 | .81 | .81 | 1.01 |
| I-K | .61 | .01 | .69 | 1.05 | 1.53 |
| J-K | 1.09 | .02 | .75 | 1.16 | 1.55 |
| J-L | .95 | .01 | .67 | .95 | 1.41 |
| K-L | .08 | .01 | .63 | .92 | 1.46 |
| K-M | .83 | .02 | .78 | .81 | 1.03 |
| L-M | 1.23 | .02 | .66 | .95 | 1.45 |
| L-N | .29 | .01 | .63 | 1.14 | 1.80 |
| M-N | -.30 | .01 | .62 | .89 | 1.44 |

To illustrate the range of within-link fit represented in Table 1, the plot for the least internally consistent link L-N is shown in Figure 3. Although there are a number of observations at or just outside the 95 percent confidence limits for this link, there are no grossly discrepant cases.

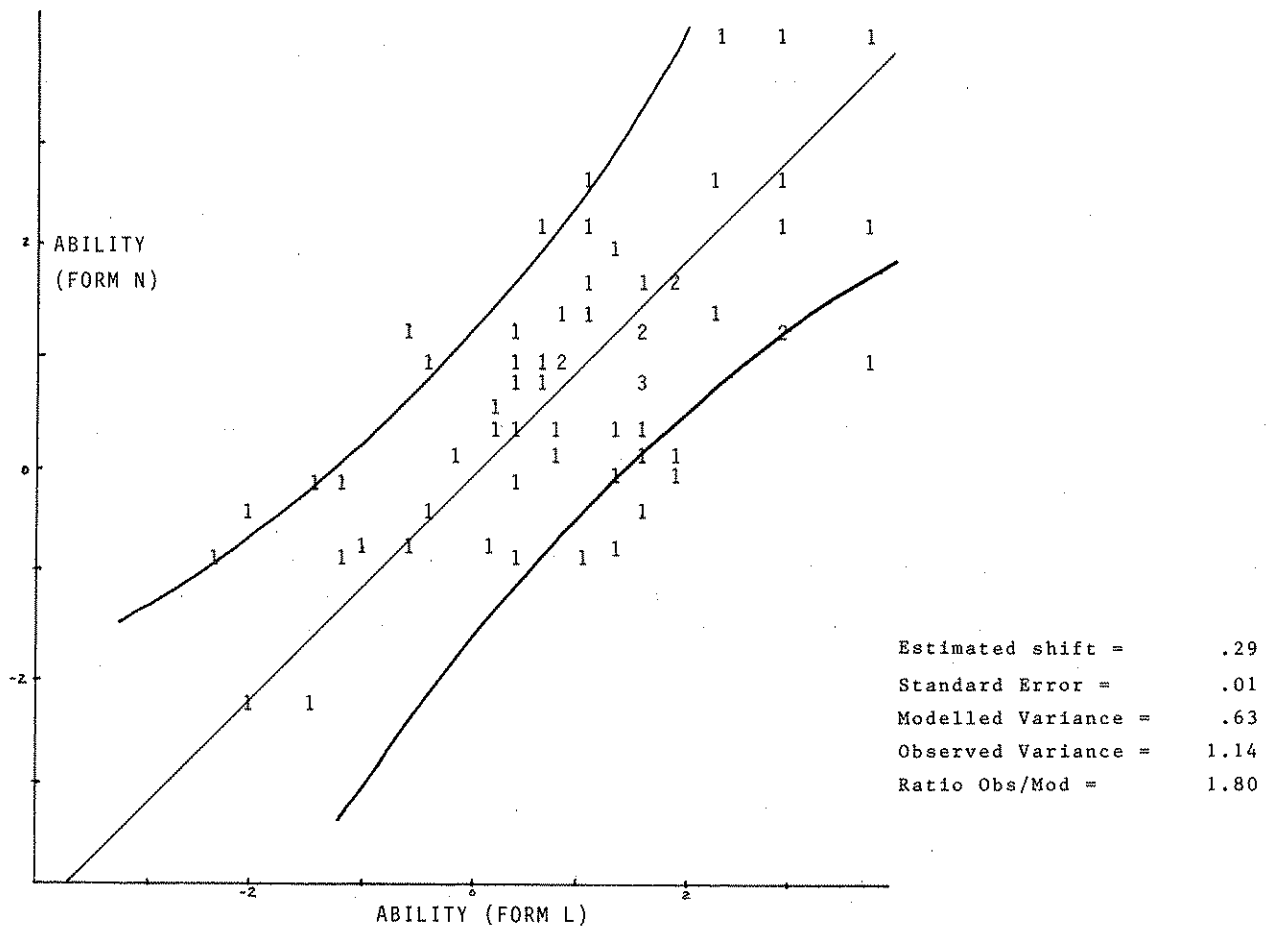


Figure 3. Link L-N.

During the construction of these tests, items which displayed poor fit to the Rasch model were either deleted or revised. This should have resulted in greater item homogeneity *within* each form. However, no statistical check on the dimensionality of items in different forms was available. The question of whether items in different forms define the same reading variable is of particular importance here because each test is based on a different piece of writing. Among these tests are stories about a mythical swamp-creature, a street-murder, and a friendly bear, and factual passages about grasshoppers, earthquakes, and parapsychology. In developing these tests we began with the proposition that each student could be usefully thought of as having *one* level of reading ability. For most students this proposition is reasonably well supported. However, as Figures 2 and 3 show, some students performed significantly better on one test than on the other. Their discrepant performances may be a reflection of their different interests in these topics, of their different approaches to different types of writing (e.g., informative vs. narrative), or of their motivation on the day of the test (the two tests were administered on consecutive days). Whatever the reason, their different levels of performance on the two tests are working against a meaningful estimate of their reading ability and warrant closer investigation.

At this point in the equating of a set of tests, the outliers in each link could be removed and the shift re-estimated (see Wright and Bell, 1981). This has not been done here. Instead, all available cases have been retained and used in the equating.

EQUATING THE TESTS

The next stage in the equating brings together the twenty-five shift estimates from Table 1 to estimate the locations of the fourteen test forms on one line of increasing difficulty. This procedure uses all the information available from this linking design about the relative difficulties of the fourteen tests. The procedure used for the equating is perhaps best explained with a small example.

At the top of Table 2 hypothetical shift estimates for a set of five test forms are shown. Notice that Form A has not been linked to either Form D or Form E, and Form B has not been linked to Form E. This is similar to the design for the fourteen reading tests in that the easier forms are not linked to the harder forms. The difference is that in the shift matrix for the reading tests there are many more missing links.

TABLE 2
THE EQUATING PROCEDURE

| FORM | FORM | | | | |
|---------|-------|-------|-------|------|------|
| | A | B | C | D | E |
| A | -- | .21 | .40 | | |
| B | -.21 | -- | .23 | 1.20 | |
| C | -.40 | -.23 | -- | 1.01 | 1.29 |
| D | | -1.20 | -1.01 | -- | .34 |
| E | | | -1.29 | -.34 | -- |
| <hr/> | | | | | |
| A | .00 | .21 | .40 | 1.41 | 1.72 |
| B | -.21 | .00 | .23 | 1.20 | 1.54 |
| C | -.40 | -.23 | .00 | 1.01 | 1.29 |
| D | -1.41 | -1.20 | -1.01 | .00 | .34 |
| E | -1.72 | -1.54 | -1.29 | -.34 | .00 |
| <hr/> | | | | | |
| T_X : | -.75 | -.55 | -.33 | .65 | .98 |
| <hr/> | | | | | |
| A | .00 | .21 | .40 | 1.40 | 1.73 |
| B | -.21 | .00 | .23 | 1.20 | 1.53 |
| C | -.40 | -.23 | .00 | 1.01 | 1.29 |
| D | -1.40 | -1.20 | -1.01 | .00 | .34 |
| E | -1.73 | -1.53 | -1.29 | -.34 | .00 |
| <hr/> | | | | | |
| T_X : | -.75 | -.55 | -.33 | .65 | .98 |

The first step in the procedure is to obtain initial estimates for the missing links. From the available links, Form B is estimated to be .21 logits harder than Form A, and Form D is estimated to be 1.20 logits harder than Form B. An initial estimate for the A-D link then is $.21 + 1.20 = 1.41$ logits. This has been inserted into the shift matrix in the middle of Table 2. Similarly, initial estimates have been obtained for missing links A-E and B-E. The mean of the five estimates in each column of this matrix is now obtained and provides the first set of form difficulty estimates T_A, T_B, \dots, T_E . Form A is estimated to be the easiest of these five forms with a difficulty $T_A = -.75$ logits, and Form E the hardest with a difficulty $T_E = .98$ logits. These five form difficulty estimates are automatically centered on zero.

These five estimates are now used to improve the estimates for the three missing links. The estimate for link A-D, for example, is calculated as $T_D - T_A = .65 - (-.75) = 1.40$. This improved estimate has been incorporated into the third version of the shift matrix at the bottom of Table 2 together with similarly improved estimates for links A-E and B-E. The five estimates in each column are again averaged to obtain a new set of form difficulty estimates. This process is continued until values of the form difficulties T_A, T_B, \dots, T_E become stable to two decimal places. In this example, because the invented shift estimates at the top of Table 2 are reasonably consistent with one another, convergence has already occurred.

This iterative procedure was carried out on the 14x14 shift matrix for the *Tests of Reading Comprehension*. On the first attempt, convergence was slow and an inspection of the shift matrix

after each iteration showed that convergence was being hampered by the shift estimate for link F-G. (This direct estimate of the relative difficulties of Forms F and G was inconsistent with the rest of the available evidence on the relative difficulties of these two forms). To achieve convergence the estimate for this link was removed from the matrix and link F-G was treated as another missing link. The final estimates of the fourteen form difficulties are shown in Figure 4.

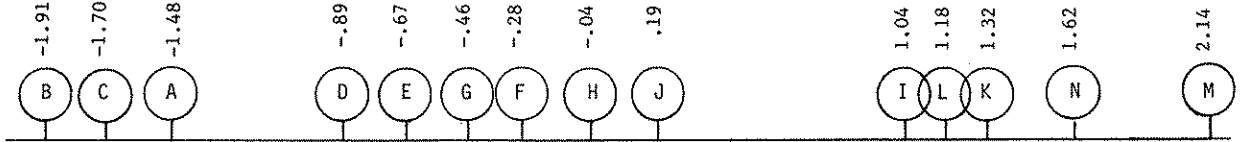


Figure 4. The form difficulties.

The fit of each link X-Y into the structure as a whole is now analyzed by comparing the shift estimate t_{XY} (based on those students who took both Form X and Form Y) with the estimate $T_Y - T_X$ of the relative difficulties of these two forms after equating. Table 3 shows the results of these comparisons. When only the group of children who took both Form A and Form B are considered, Form B is estimated to be .29 logits easier than Form A. However, when all the information available about the relative difficulties of Forms A and B is considered, Form B is estimated to be .43 logits easier than Form A, a difference $t_{XY} - (T_Y - T_X)$ of .14 logits.

TABLE 3
BETWEEN-LINK ANALYSIS

| LINK | SHIFT ESTIMATE | | | LINK | SHIFT ESTIMATE | | |
|------|----------------|-------------|------|------|----------------|-------------|------|
| | t_{XY} | $T_Y - T_X$ | DIFF | | t_{XY} | $T_Y - T_X$ | DIFF |
| A-B | -.29 | -.43 | .14 | G-I | 1.42 | 1.50 | -.08 |
| A-C | -.29 | -.22 | -.07 | H-I | 1.09 | 1.08 | .01 |
| B-C | .18 | .21 | -.03 | H-J | .10 | .23 | -.13 |
| B-D | 1.25 | 1.02 | .23 | I-J | -.89 | -.85 | -.04 |
| C-D | .83 | .81 | .02 | I-K | .61 | .28 | .33 |
| C-E | .91 | 1.03 | -.12 | J-K | 1.09 | 1.13 | -.04 |
| D-E | .28 | .22 | .06 | J-L | .95 | .99 | -.04 |
| D-F | .61 | .61 | .00 | K-L | .08 | -.14 | .22 |
| E-F | .49 | .39 | .10 | K-M | .83 | .82 | .01 |
| E-G | .08 | .21 | -.13 | L-M | 1.23 | .96 | .27 |
| F-G | .39 | -.18 | .57* | L-N | .29 | .44 | -.15 |
| F-H | .33 | .24 | .09 | M-N | -.30 | -.52 | .22 |
| G-H | .39 | .42 | -.03 | | | | |

The form difficulty estimates for these fourteen tests cover more than four logits of this reading variable. Nevertheless, more than one half of the original shift estimates differ from the final bank estimates by less than one-tenth of a logit. The poorest link in terms of its consistency with the surrounding links is link F-G. This was the link that had to be removed to facilitate convergence. Based on the performances of the group of sixth grade children who took Forms F and G, Form F is estimated to be .39 logits easier than Form G. Based on the other evidence concerning the relative difficulties of these two forms, Form F is estimated to be .18 logits harder than Form G, a difference of .57 logits! While it is not uncommon to encounter a few inconsistent links in a bank such as this, their presence lowers the coherence of the bank as a whole.

In this case, an interesting hypothesis is available for the misfit of link F-G. Upon completion of Form F, one group of sixth grade children reported that they already knew this story about the swamp-creature. One girl said she had seen it on television earlier in the year. Another pointed out that the book was in the school library. If some sixth grade children were already familiar with this story, then they may have performed unexpectedly well on Form F and so made it appear easier rather than harder than Form G. This appears to have been the only one of these fourteen passages with which some children already had some familiarity.

DISCUSSION

This has been an exploratory study of the effectiveness of using groups of common-persons to vertically equate a set of tests. The attempt to equate these fourteen tests is based on the proposition that they all define the same reading variable--a proposition that could be questioned in view of the wide range of content, style and difficulty represented in these fourteen passages of writing. Nevertheless, for most of these students this proposition appears to have been reasonably well supported. For those students with two significantly different ability estimates, an opportunity exists to learn something about the way in which their ability to get meaning from text interacts with the nature of the passage.

While most of the links in this test bank are less internally consistent than has been modelled, an inspection of even the worst link (Figure 3) suggests that, for practical purposes, this level of misfit may not be intolerable. (In examining Figures 2 and 3 it must be remembered that each of the ability estimates in these links is based on about twenty items. In contrast, difficulty estimates in common-item links are usually based on several hundred persons). The utility of these common-person links is further supported by the between-link analysis.

REFERENCES

- Divgi, D.R. Model-free evaluation of equating and scaling. Applied Psychological Measurement, 1981, 5, 203-208.
- 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.
- Gustafsson, J-E. The Rasch model in vertical equating of tests: A critique of Slinde and Linn. Journal of Educational Measurement, 1979, 16, 153-158.
- Holmes, S.E. Unidimensionality and vertical equating with the Rasch model. Journal of Educational Measurement, 1982, 19, 139-147.
- Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale NJ: Erlbaum, 1980.
- Loyd, B.H. & Hoover, H.D. Vertical equating using the Rasch model. Journal of Educational Measurement, 1980, 17, 179-193.
- Mossenson, L.T. & Masters, G.N. The Tests of Reading Comprehension: User's Manual. Perth: Education Department of Western Australia, 1983.
- Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960 (University of Chicago Press, 1980).
- Rentz, R.R. & Bashaw, W.L. Equating reading tests with the Rasch model. Athens, Georgia: Educational Resource Laboratory, 1975.
- 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. The Rasch model, objective measurement, equating and robustness. Applied Psychological Measurement, 1979a, 3, 437-452.
- Slinde, J.A. & Linn, R.L. A note on vertical equating via the Rasch model for groups of quite different ability. Journal of Educational Measurement, 1979b, 16, 159-165.
- Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14,
- Wright, B.D. & Bell, S.R. Fair and useful testing with item banks. Research Memorandum No. 32, MESA Psychometrics Laboratory, Department of Education, University of Chicago, 1981.
- Wright, B.D. & Masters, G.N. Rating scale analysis. Chicago: MESA Press, 1982.