SECTION FIVE

THE STATISTICAL MODEL FOR TEST EQUATING

The model recommended for test equating is the general Rasch measurement model. There are two basic reasons for recommending the model. The first concerns the efficiency of the model in providing equivalences among test forms in which not every person attempts every test form; the second concerns the diagnostic opportunities the model provides for understanding the data - in particular, the possibility of identifying and making quantitative equivalences which are meaningful, and identifying distinctions in kind when quantitative differences are not meaningful. These properties are now described in greater detail.

5.1 Efficiency of the model in providing equivalences among test forms

5.1.1 Different kinds of items

The same model can be readily applied simultaneously to items that are dichotomously scored (for example, but not exclusively, multiple-choice items) and those that are in the form of graded responses or partial credit, and even to tests as a whole.

5.1.2 Distribution-free property of the estimates

The model provides the relative difficulty estimates (and other parameter estimates) of the items which are on the same linear scale, and these estimates are independent of the distribution of the actual achievements of the people tested. It is necessary that the persons and the items are reasonably well targeted to each other. This means that the test scores of persons should be relatively evenly spread out across the scale without too many extreme scores, neither at the maximum nor at the minimum end of the scale. However, no assumption, such as that the distribution of persons is normal, needs to be made. The distributions of the persons is an empirical question which is answered from the data available from the individual teaching programs of the States/Territories and which are equated through this procedure.

5.1.3 Linking of test forms and items

It is not necessary for all persons to respond to all items in order for the item parameter estimates to be obtained. It is necessary that there is sufficient overlap among the combinations of persons and items to make the estimates stable, and

this can be ensured by the sampling design. Conceptually, and in some computer programs in practice, the responses of all people who have completed a pair of items are used to obtain the relative difficulty (and other parameters) of that pair of items. Not all people have to complete each pair of items. Therefore, different items can be compared with different combinations of other items. These pairwise comparisons among items across different groups of persons are carried out simultaneously, and this provides item parameter estimates which are on the same relative scale. In this way, the items and test forms are said to be linked.

There are curently test equating initiatives among different groups of States/Territories using the Rasch measurement model. These, however, do not provide country-wide linking as proposed for Phase Two to enable nationally comparable results.

5.1.4 The achievement continuum

The location of the items and the persons on the same linear achievement continuum also makes the capabilities of the students tangible.

5.1.5 Equating tests directly

The process of equating of *items* in pairs can be generalised to comparing *tests* in pairs. This type of comparison can be instructive in appreciating the principle of the equating. For example, suppose two tests are to be equated. Then the scores of all persons who have taken both tests are considered. In the Rasch models, the key statistic is the total score that each person has on the two tests. Given the total score on the two tests, the key information is that total score composed, and in particular, are the scores on one test consistently higher than on another test? For example, suppose there are two tests where each has a maximum score of 25. Then the total maximum score is 50, and the minimum score is 0. If the two tests are identical in difficulty at all levels, then it would be expected that on the average, a total score is composed of approximately equal scores on the two tests - for example, a total score of 30 would be composed of approximately 15 from each test. Figure 5.1 shows a hypothetical example of a set of scores on a Test 1 relative to the total score on Test 1 and Test 2. It is evident that Test 1 is easier than Test 2 because the scores on Test 1 are on the average higher than the scores on Test 2 for the same total score. The estimation of the model parameters is consistent with the graphical display in figure 5.1.

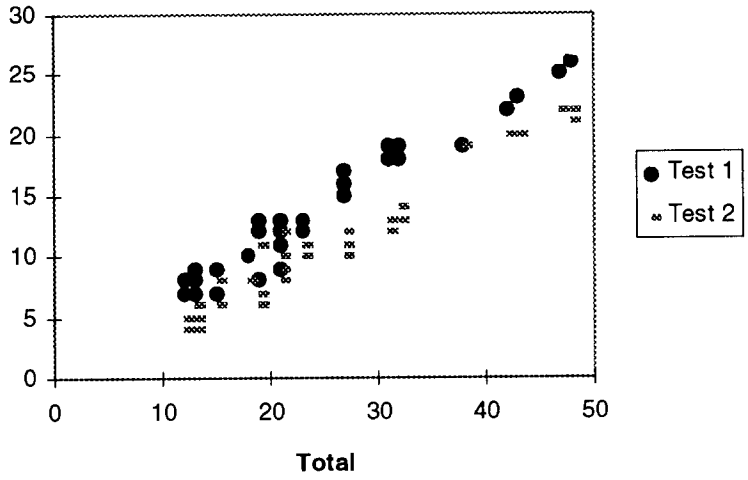

Figure 5.1 Equating two tests directly

5.1.6 Equating tests through item difficulties

Once the item parameters have been estimated, an estimate of each person's achievement on the hypothesised continuum of achievement, and which is on the same scale, can be obtained for all scores on any subset of items. This means that the items from specific tests can be considered as a subtest from the total set of items, and the scores of these subtests automatically equated to each other. For example, a score of 20 on one test which may have a maximum score of 25 might be equated to a score of 32 on another test which may have a maximum score of 40. Figure 5.2 shows such an equating for two subtests of 10 mutually exclusive items from a test of 20 items. It is also evident from figure 5.2 that in everyday terms, Subtest 1 is easier than Subtest 2 because for the same total score of 6, the achievement from Subtest 1 (0.10) is less than the achievement on Subtest 2 (0.76).

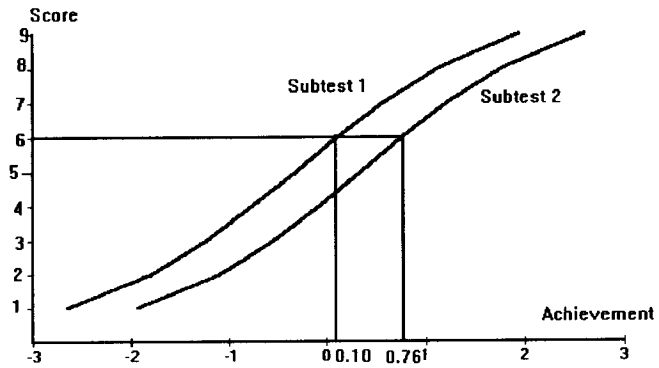

Figure 5.2 Equating two tests through the item parameters which are on the same scale

5.2 Diagnostic opportunities the model provides for understanding the data

It is generally understood that the general Rasch model is unidimensional. Because there is sometimes confusion in appreciating this feature of the model and its full implications for understanding the data, this is now considered in some detail. In order to do so, the equation of the model is presented. The key point in presenting the equation is to identify the variables in the model and to facilitate the general discussion by using these variables, rather than to discuss the form of the model.

5.2.1 The equation of the model

The equation of the model in the way it is generally represented takes one of the following forms:

$$\Pr\{X_{ni} = x\} = \frac{1}{\gamma_{ni}} \exp\{-\tau_{1i} - \tau_{2i} \ldots - \tau_{xi} + x(\beta_n - \delta_i)\}$$

(Equation 5.2.1a)

where

$$\gamma_{ni} = \sum_{x=0}^{m} \exp\{-\tau_{1i} - \tau_{2i} \ldots - \tau_{xi} + x(\beta_n - \delta_i)\}$$

or its equivalent

$$\Pr\{X_{ni} = x\} = \frac{1}{\gamma_{ni}} \exp\{-\delta_{1i} - \delta_{2i} \ldots - \delta_{xi} + x\beta_n\}$$

(Equation 5.2.1b)

where $\delta_{xi} = \tau_{xi} + \delta_i$ and $\sum_{x=1}^{m_i} \tau_{xi} = 0$, and where

- β_n is the location of person n on a hypothesised linear continuum of achievement
- δ_i is the location of item i on the same hypothesised continuum, generally termed the difficulty
- τ_{xi} are a series of thresholds on the latent continuum between the categories in a graded response and become the one item difficulty in the case of the dichotomous response item.

5.2.2 Unidimensionality is relative

It is evident that there is only the one parameter β_n which characterises the person. This implies that there is only the one kind of achievement that characterises all people, but that this might vary from person to person. For this reason, it is considered that the model is *unidimensional*, that there is only one dimension which governs the responses of the persons to the items. However, an area of study such as literacy has many components, and the issue that is often raised is the how the hypothesis or assumption of unidimensionality can accommodate such a complex subject matter.

Understanding this issue is best approached by recognising that unidimensionality is a *relative*, not an absolute, matter. For example, at a very broad level of dimensionality, school performance across a range of subjects can be summarised into a single concept (even if it is not formally quantified) to indicate whether a person is performing relatively well at school; at a finer level of dimensionality, performance within the different subject disciplines can be summarised into a single concept to indicate whether a person is performing relatively well in each subject at school; at an even finer level of dimensionality, performances within the different components of a subject discipline can be summarised into single concepts to indicate whether a person is performing relatively well in each of the components within that discipline, and these components can themselves be divided further, and generally are in the development of teaching programs.

In the primary analysis of the responses of persons to items according to equations 5.2.1a or 5.2.1b above, it is hypothesised that there is a single location on the achievement continuum which governs the responses of the person to all items. All items and persons are located on the same achievement continuum with a point estimate and a standard error of each estimate.

5.2.3 Diagnostic analysis

Following the primary analysis, a secondary analysis which identifies the degree to which the hypothesis of a single achievement continuum can be sustained needs to be carried out. Given the variation among the components of literacy, and the different emphases given to the different aspects of literacy that have been identified, it is unlikely that even at a general level the data will be sufficiently unidimensional that simple quantitative comparisons of differences in degree of performance would be sufficient. In this case the analysis has features of scaling as well as equating. *Scaling* is the term generally used to decribe the process of making equivalent, in terms of difficulty, the scores on tests that are inherently different in content, while *equating* is used to describe the process of making equivalent the scores on tests that are inherently similar in content.

This secondary analysis is often called a *fit* analysis, and existing software provides various kinds of checks between the hypothesis of the single achievement according to the model and the data. Although these general diagnostics are important, often in major studies it is necessary to consider specialised analyses to identify the degree to which a single achievement continuum can be accepted and where more detailed interpretations are required. Other complementary analyses can also be carried out at this level.

To illustrate the point, two kinds of analyses will be considered, and there are many other possibilities.

(a) An empirical common core

From the analysis, it is possible to diagnose which subset of items seems to conform most closely to the hypothesis of a single dimension, and to complement this with a content analysis. It is possible that an empirical common core of studies across States/Territories will emerge.

Simultaneously, and complementary to any emergence of a common core among States/Territories, the analysis can diagnose which elements of the curriculum are not common and are emphasised in each State/Territory. Whether such a common core emerges or not is an empirical matter, and cannot be decided in advance, but it is important to appreciate the kind of analysis that can be carried out to identify if it exists to any useful degree.

(b) Comparisons in performances on tests

It is recommended that each person in each State/Territory complete a Home Test and a Foreign Test. If the hypothesis of a single achievement continuum is acceptable, then each person's achievement level estimated from the Home Test should be statistically equivalent to the achievement level estimated from the Foreign Test. Any statistical differences would indicate that there is a different level of achievement on the Home Test from the Foreign Test. Such a difference would not necessarily preclude making more global comparisons at a broader level of dimensionality. For example, if each State/Territory achieved somewhat higher on the Home Test compared to the Foreign Test, then it might still be valid to take the achievement on both tests as a kind of average performance, and to make relevant comparisons.

Diagnostic analyses might be carried out at the level of the different components of literacy, for example, reading and writing where the data are available, and even at a finer level by studying the performances on individual items. Such analyses, and the decisions at which level and to what degree a quantitative comparison is meaningful, and at which level and to what degree a difference in kind is meaningful, requires sound understanding of item response theory in general, and Rasch models for measurement in particular.

The above commentary has focused on making the test scores equivalent. With equivalent scores, the data from the States/Territories can be analysed. Explanatory variables and multilevel structuring of the data (student, school and State/Territory) may be incorporated in some of the analyses. This may give an additional understanding of the data beyond comparisons of distribution of the scores, and may be relevant to policy makers. However, this kind of analysis will depend in part on the data themselves. For example, it is possible to invoke directly more sophisticated analyses of a single dependent variable for any common core which emerges, and to understand the factors that explain variation in the scores across States/Territories.

5.2.4 Summary

In Phase One, the plan for a statistical model that enables the establishment of equivalences across tests of aspects of literacy used in States/Territories is proposed. In Phase Two, the following steps are to be taken within the two major classes of analyses, one for forming the equivalences, the other for analysing the State/Territory level data.

Forming equivalences

- Students in different States/Territories take two tests from the existing State/Territory tests.
- The two tests taken by students in each State/Territory are in different pairs from State/Territory to State/Territory, providing an overlap among tests and students from the States/Territories.
- The total number of distinct tests is six.
- The items in the six test forms are then considered as one large set of items.
- A person (rows) by items (columns) matrix is formed with the score of each person on each item that the person attempted. There will be blanks (missing data) where persons have not taken a particular item. Missing data can be structural, as in the case where the test forms are not administered to students, and this will be substantial, or non-structural, as where a person is administered the test but the student has skipped or not reached the item. In the case of forming equivalences, non-structural missing data can be considered simply as missing, or students with any missing response can be eliminated (providing it does not reduce the database substantially). The analyses should be compared for stability, there being some evidence that the latter may be more stable.
- The matrix of person by item responses is submitted to a single analysis according to the general Rasch model. This is possible because of the overlap among persons and items, even though not all persons have taken all items. (Other complementary analyses may be carried out on this data matrix as well, for example, that based on correlations.)
- The single Rasch analysis gives a location (difficulty) value for each item on the continuum. These locations place all items from all tests on the same scale or metric.
- Each score on each of the original tests is then transformed to a score on the same scale.
- The quality control of these transformations is provided through the diagnostic analyses within the Rasch model (and any complementary analyses). Some different subsets of items may be need to be treated as separate dimensions and the analysis in the previous step repeated.
- Location of National Profiles are located on the common scales from the previous two steps.

Analysing the State/Territory level data

- State/Territory level data must be available in Phase Two.
- Any non-structural missing data must be treated simply as missing data so that all students who have been tested provide data for the comparisons.
- Within- and between-State/Territory differences on the dimensions identified and on the State/Territory level data must be examined, recognising where different States/Territories have responded to different test forms in any substantive interpretation.

- Having distributions of scores displayed on the common metric of the tests as a whole and of the subsets of items that are identified as separate dimensions.
- Where possible, analyses (for example, hierarchical modelling) which will explain the variation in performances at student, school, and State/Territory levels are carried out.