Appendix 11- Summary of equating approaches and issues

This appendix briefly summarises four approaches to equating. All four are used at various points in the analysis in Chapter 8. Traditional equating approaches include Mean, Linear and Equi-percentile equating (Kolen & Brennan, 2004; von Davier, Holland, & Thayer, 2004). A fourth process using independent Rasch scaling of the test and teacher scales and then equating the means and SDs, a Rasch scaled Linear equating, is also applied.

Mean equating

This equating process is the simplest, transposing the individual data points so that the mean of one scale equals the mean of the other. Generally the process is inadequate for effective transformation of scales, except in the rare case of equal SDs, as it takes no account of the difference in the spread of the two scales. It is mentioned briefly as it is used at the end of Chapter 8 in Figure 8.16 and subsequent figures to 'equate' scores to compare the model test data developed in Chapter 6 to the full Years 1 to 8 teacher data described in Chapter 7. It is used also as a process in this analysis to 'equate' the trajectories of the scales.

Linear equating

Linear equating transforms both the mean and the spread of one data set to match the other. The expression to do this is a simple linear transformation, and thus is insensitive to any non-linear relationship between the scales. Linear transformation is used in the Rasch equating described below with the teacher scale initially converted to equal interval units by the Rasch model analysis.

Equi-percentile equating.

The equi-percentile equating approach establishes the score at a number of percentile points on each scale and assumes that the scores from both scales at these points are equivalent. The relationship of one scale to the other need not be linear.

Rasch scaled linear equating

A Rasch model analysis of the full teacher assessment data set (not just the matched cases) is conducted and then the logit scores for the set of common students equated by the linear method. This brings the mean and SD of the set of teacher assessed students common to the test, to the test mean and SD. Test scores are not used as anchors for reasons discussed below

The transformation formula found for common persons is then applied over all teacher assessments to bring all teacher-assessed cases to a test score equivalent for all cases across all year levels. An estimate of model measurement error for each student is derived in the Rasch model analysis. This error from the profile level value to logit translation is one of the error factors additional to the possible variation in judgement skill and scale calibration of teachers.

Further issues in the equating process

Two additional issues arise in the consideration of the equating of the teacher and test scales: the timing of the assessments and the different levels of performance of a skill from a teacher or test perspective.

For the 1997 data there is an issue of timing. The students were tested in early August while teacher assessments extended over a period from early October to mid November. As a result the estimated learning status for students in October will, on average, be higher than at the point of testing. Based on the analyses in Chapter 7, where a relationship of learning with age based on teacher judgements is established, the real learning status at the earlier testing point would be about 0.1 profile units lower than that recorded by teachers. No direct adjustment is made for the time shift in this analysis. There is also a possibility that teacher judgement assessments were influenced by test results arriving before the teacher assessments were made. Based on the view taken by the teachers about test results and the lack of linking of the test scale to the SPFAS scales, the likelihood of test results directly influencing teacher judgement assessments is very low.

Teacher judgement assessment data are treated, for the purposes of comparison, as if they occurred in August with a consequence that the equating arrangements will over estimate the relationship of profiles level to test scale units. Had the collection of data been repeated in subsequent years some form of adjustment would have been required. The data for 1998 were both collected in August removing the timing and influence problems.

The second issue that applies to both collection years is the potential difference in the criterion that teachers apply to a judgement of learning compared with the criterion in the test Rasch analysis model. In the Rasch model, as applied to the test data, a student is placed at a point where the odds of success on an item are estimated to be 50:50. The item scale itself is based on items positioned at the points where "the difficulty ... of an item … is the point on the latent variable (uni-dimensional continuum) at which the highest and lowest category have equal probability of being observed. For a dichotomous item, this is the point at which each category has a 50% probability of being observed" (Linacre, 2006, p. 300). This is the situation that applies for the test.

For the teacher judgement it is unlikely that a teacher will regard a level of performance of some behaviour as being achieved if the student can only perform it half of the time. Thus the equating process is for a teacher judgment at a higher criterion level than the test, for the same target behaviour. For the purpose of equating, this criterion shift, assuming it is relatively consistent across teacher assessments, will make no difference to general equating. It will however have implications in the interpretation of the relationship. Effectively the teacher scale will be displaced relative to the test scale, when performance of an actual skill is observed. Based on Masters et al. (1990) documenting the initial design of the Basic Skills Testing process for NSW, this concern is addressed in the conversion of students' test scores into Bands and the presentation of item difficulties in item maps, by rescaling the items to 0.7 probability of success rather than 0.5. However, when dealing with individual student data in Kidmaps (individual progress maps) the items are reported at their p=0.5 level. The data in this analysis are considered at the p=0.5 level. Based on the analysis in Chapter 6 the data analysed in this thesis were created at p=0.5.

As a result, taking the case of a specific behaviour, the test process will estimate it as 'achieved' well before the teacher. Since a focus on individual skills or behaviours is not considered directly in this analysis, this displacement will not effect teacher and test comparisons, which will be equated as if the difference between the scales does not exist. Further refinement of teacher judgements however would need to consider the practical implications. It also raises another source of variation in teacher judgement assessments. Teachers could all be generally aligned to the test scale, in principle, but apply idiosyncratic performance criteria, adding to the variability in aggregated teacher judgements.