Chapter 8 Teacher and test assessment compared

> … such is the hegemony of traditional psychometrics, that these alternative assessment systems are widely characterised as 'soft' and 'unreliable'. The pioneering work of our best teachers has run far ahead of the available theory, and I believe the lack of theoretical support from the academic community for these innovative practices has made it much easier for politicians to deride and dismiss any assessment practice that does not meet their own aims.
>
> Wiliam, 1994, p. 17-18.

This chapter applies a range of methods to convert teacher judgement assessments to the scale of test assessments so that teacher / test comparisons can be made. It is already clear from the different shapes of the trajectories of the IRT based test view (Chapter 6) and the profiles based teacher view (Chapter 7) that the conversion of one assessment process to the scale of the other cannot be a simple linear transformation. The purpose in comparing the two assessment processes is to establish the degree to which they were interchangeable in 1997 and 1998. The chapter considers a variety of methods for equating the scales of the two processes. Once the scales are approximately equated it should be possible to understand the degree to which judgements made by teachers can be considered as equivalent to the scores provided by tests. The validity of the assumption that teacher and test assessments can be compared is also considered.

Equating Teacher and Test scales

Assumptions

Prior to addressing the process for bringing the two assessment arrangements to a common scale, some discussion is required of the validity of the assumption that the two processes are addressing the same dimensions in each case (English compared with Literacy, Mathematics compared with Numeracy). The equating of the teacher judgement and test scales within specific learning areas is logical only if the case can be established that both processes are assessing the same dimensions of learning. The broadness of the English and mathematics traits and the commonsense understanding of these, imply a strong likelihood that the order, at least, of students on each scale could be expected to be similar whether teacher or test assessment is used. As part justification for continuing the equating processes, the correlations of the scores on the two scales are 0.659 (n=1275) for English/Literacy and 0.57 (n=2105) for Mathematics/Numeracy (see Tables 8.1 and 8.2 later). These are lower correlations than are expected in parallel test forms, but indicate a reasonable degree of consistency of order from the two assessment processes.

The assumptions that the English learning area and Literacy tap the same underlying latent trait of learning or that the Mathematics Learning area and Numeracy are similarly derived, are not unique to this study. As described in Chapter 4, the Victorian student assessment system has operated on this assumption in very similar circumstances. However a much closer alignment of the test to the common curriculum framework applies there as both the test and teacher judgement assessments relate directly to the VELS/CSF frameworks. The teacher judgement assessment frameworks were very similar in SA and Victoria in 1997 and 1998; and, based on the author's observation, the test items, while developed to different specifications, appear to be similar in style.

Based on the precedent actually operating in Victoria, there is justification for assuming that the underlying latent dimensions of the tests and those for teacher judgement assessments are similar enough to explore converting data from both sources to a common scale.

Data sets used

The initial step in the equating of the teacher profile level scales with the test scales is the matching of records from the test files with records from the teacher assessment files to find students in common. Using the student identifier used in Chapters 6 and 7 to assign dates of birth, the author compared the records in the teacher-assessed file with the test files (student identifiers now added) to find cases common to both files. As described in Chapter 6 only 72% of test cases were assigned a date of birth for 1997 and 75% to 80% for 1998, depending on the test. Of the sample of students assigned teacher judgement assessments, 64% of cases in the combined Year 3 and Year 5 set for 1997 were matched (n=1275), and 68% for 1998 (n=2105) (Table 8.3). The proportions of the test cohorts who were also assessed by teachers using the SPFAS approach were 5% in 1997 and 9% in 1998. The matched cases may not be randomly selected from their parent distributions.

The teacher assessments

The strand scales used by teachers have eight levels of development from the beginning of school to Year 12. Ten scale positions within each of the eight levels, obtained from teachers clicking on a progress line, were used as progress indicators within a level. As a result each strand has a range of 90 score points from 0 to 8.9. Scale positions used for Years 1 to 8 only are in the range 0 to almost 7.0, that is approximately 70 scale positions are used across these 8 Year levels. For the Year levels 3 and 5 only, most assessments are in the range 1.0 to 5.0, although the full range is 0.2 to 5.9.

In the English learning area, the correlation of Speaking and Listening with the test at Year 3 is lower than for the other two strands (Table 8.1). However the correlations of Speaking and

Listening with the other two teacher-assessed strands are high (0.92 and 0.93). These between teacher judgement strand correlation values are not reported in Table 8.1.

Speaking and Listening was not included in the averaged scores analysed in Chapter 7 on the grounds that the tests, with which the teacher judgements were to be ultimately compared, covered only Reading and Language, the latter in written forms only. That position is modified in this chapter. All three strands are used in the Rasch analysis to make the analysis feasible. Using three strands allows three items for each student.

The correlations of the strands with the test at each year level separately and as a combined data set are shown in Tables 8.2

Table 8.1 Correlations of English teacher assessments with Literacy test assessments – 1997

N=1275	Teacher assessed- Reading	Writing	Speaking & Listening	Correlation with the average of all three strands
Test Year 3-Literacy	0.55	0.52	0.39	**0.52**
Test Year 5-Literacy	0.60	0.59	0.51	**0.60**
Test Combined Years	0.67	0.65	0.58	**0.66**
Female –combined	0.67	0.66	0.58	**0.66**
Male -combined	0.65	0.62	0.56	**0.64**

In the mathematics learning area the Working Mathematically strand correlates between 0.75 and 0.78 with the other four mathematics strands , while they correlate with each other in the range 0.85 to 0.9. It is assumed that Working Mathematically is either measuring a different dimension to some extent or was less well understood by teachers. Either way Working Mathematically appears to be different to the other strands. As a result the Working Mathematically strand is not included in the averaged score for each student in the analysis in this Chapter, nor as an item in the Rasch analysis described later. The test-teacher correlations by strand are shown in Table 8.2.

Table 8.2 Correlations of Mathematics teacher assessments with Numeracy test assessments-1998

N=2105	Teacher assessed- Chance	Measure-ment	Number	Space	Correlation with the average of all four strands	(Working Mathematic-ally)
Test Year 3-Numeracy	0.37	0.40	0.42	0.39	**0.42**	(0.32)
Test Year 5-Numeracy	0.48	0.46	0.49	0.47	**0.50**	(0.37)
Test Combined Years	0.53	0.55	0.57	0.54	**0.57**	(0.47)
Female –combined	0.51	0.52	0.54	0.52	**0.55**	
Male -combined	0.56	0.57	0.59	0.56	**0.59**	

The correlation coefficient for the combined data set of the test and teacher assessments for English/Literacy is 0.659. For Mathematics/Numeracy the correlation of the combined data set teacher with test assessments is 0.571. These values are less than usually expected for parallel forms of test-based assessments.

Comparing raw scores

A range of statistical characteristics of the Year 3 and Year 5 cases with both a teacher and test assessment are reported in Table 8.3.

Table 8.3 General Statistical Characteristics of common cases of Teacher assessments and Test assessments, 1997 and 1998

		1997 Teacher (English) Profile units	Test (Literacy) Logits	1998 Teacher (Mathematics) Profile units	Test (Numeracy) Logits
Year 3	**Mean**	2.20	0.50*	2.13**	0.17
	Median	2.20	0.61	2.13	0.22
	SD	0.55	1.26	0.55	1.41
	SE (Mean)	0.02	0.05	0.02	0.04
	Min	0.17	-6.42	0.50	-6.18
	Max	4.00	3.80	4.65	4.58
	Skewness	0.09	-0.24	0.06	-0.24
	Kurtosis	3.39	3.97	3.14	5.46
	N	702	702	1035	1035
Year 5	**Mean**	2.98	1.74**	2.96**	1.35
	Median	3.00	1.78	2.95	1.38
	SD	0.67	1.19	0.71	1.22
	SE (Mean)	0.03	0.05	0.02	0.04
	Min	1.13	-2.06	0.60	-5.40
	Max	4.97	5.36	5.90	5.96
	Skewness	0.14	-0.21	0.25	-0.68
	Kurtosis	3.44	3.44	3.61	8.18
	N	573	573	1070	1070
Combined	**Mean**	2.55	1.06	2.55	0.77
	Median	2.50	1.07	2.50	0.83
	SD	0.72	1.37	0.76	1.44
	SE (Mean)	0.02	0.04	0.02	0.03
	Min	0.17	-6.42	0.50	-6.18
	Max	4.97	5.36	5.90	5.96
	Skewness	0.33	-0.21	0.40	-0.46
	Kurtosis	3.24	3.45	3.37	5.27
	N	1275	1275	2105	2105
Year 3 assessments (matched and total)		702/1005 (70% matched)	702/12437 (6% matched)	1035/1537 (67% matched)	1035/12794 (8% matched)
Year 5 assessments (matched and total)		573/996 (58% matched)	573/11973 (5% matched)	1070/1541 (69% matched)	1070/12471 (9% matched)
All cases (Year 3 + Year 5)		1275/2001 (64% matched)	1275/24410 (5% matched)	2105/3078 (68% matched)	2105/25265 (9% matched)

* Difference from means in Tables 6.5, 6.9, 7.1 or 7.2 significant at 5% level based on t-test

** Difference from means in Tables 6.5, 6.9, 7.1 or 7.2 significant at 1% level based on t-test

The table indicates the general statistical characteristics of the students common to both assessment processes. Listed are teacher assessed profile level values (based on the averaging

of strands for each student) and test values as logit scores obtained from the item linked vertical scale for the tests. The general statistics for the full test populations are found in Tables 6.5 (1997 test) and 6.9 (1998 test) and in Tables 7.1 and 7.2 for the teacher assessments.

The means and medians in each column of Table 8.3 are close to those in each of the 8 separate Year level samples and the 4 combined samples for all but the 1997 Year 3 Literacy test where they differ by 0.09 of a logit. The skewness and kurtosis statistics are generally comparable to the original data sets in Tables 6.5, 6.9, 7.1 and 7.2. The test means for 1998 are not significantly different from the full cohort but the means for 1997 differ by up to 0.2 logits, and are different beyond the 5% and/or 1% significance levels. The means of the sub-samples with both teacher assessments and test assessment compared to the original full sample teacher means for 1997 are not significantly different. For 1998 the sub-samples with teacher-and test assessments have means approximately 0.1 of a profile level above the full sample means. The SDs in the sub-samples of teacher assessments with matched test cases are comparable to those for the full test cohort in Tables 6.5, 6.9, 7.1 and 7.2.

That the students with both test and teacher assessments do not have means identical to their parent samples i.e. the selection may not be random, is not regarded as critical. The purpose in this first stage is to examine the assessment scores only for the set with both assessments, in an attempt to equate the teacher judgement assessment scale to the test scale. Necessary for such a process is a good spread of cases on both the test scale and the teacher scale. The common cases provide this spread.

The assessments are from multiple teachers. The set of records where a test and teacher assessment can be compared depend in the first instance on the allocation of the student identification codes to each file. For various reasons not all students who were assessed by both processes could be identified. As a result some teachers who provided assessments may have been removed, or at least part removed, from this part of the analysis. While it is impossible to know which teachers were affected, an unintentional bias in the deletions might have occurred. However, for the purpose of the analysis the failure to allocate student identifiers is assumed to be random, or, at least, trivial. Two further issues arise that are important in the appropriateness of equating the scales. These are addressed in detail in Appendix 11. For 1997 the data for teachers were obtained up to three months later than the test assessments. It has been assumed for equating purposes that this time difference does not exist. As the process is to establish a relationship for a one-off analysis, this time difference is immaterial to the results. In reality the conversion relationship should be set such that each profile value takes a slightly lower test scale value (of the order of 0.1 logits lower). For the 1998 data, tests and teacher judgement assessments were at almost the same time. The second

issue is the difference in the criterion used by the test (50:50 odds) and that likely to be used in teacher judgement assessments where mastery of a skill is required. For the teacher mastery of a behaviour or skill might require expression at 80% to 90% of the time. This important difference is not adjusted for in the analysis and is discussed in more detail in Appendix 11.

The nature of the assessments as continuous or discrete data also needs comment. Teacher judgement assessments can be considered as approximately continuous through the averaging process applied across strands. Test data are scaled on a Rasch scale at points expressed to two decimal places. Even though the actual data points bear a direct relationship to the original number correct scores (i.e., they are discrete), they are assumed to be continuous for the purpose of the explorations. The concentration of the points on the test scale as shown in scatter diagrams, highlight that the test data points are not continuous in practice (Figures 8.8 and 8.9).

Equating approaches

The principles of the range of equating processes used are summarised in Appendix 11. The processes described include mean, linear and equi-percentile equating. All are used in sections of the analysis. A specific linear equating approach (non-anchored Rasch scaled linear equating) is used as one basis to equate the teacher and test scales. In this process the two scales are developed independently for all the teacher and test cases using the Rasch model. Then for the common students the means and SDs are equated. Based on the summaries by Year level for the teacher-assessed cases presented in Chapter 7, the relationship of mean learning status with Year level appears to be linear. From a test perspective the trajectories of mean learning status are curved, with growth increments diminishing with Year level (Chapter 6). As a result, the arrangement to convert teacher assessment scores into the framework of the test over the full range cannot be linear.

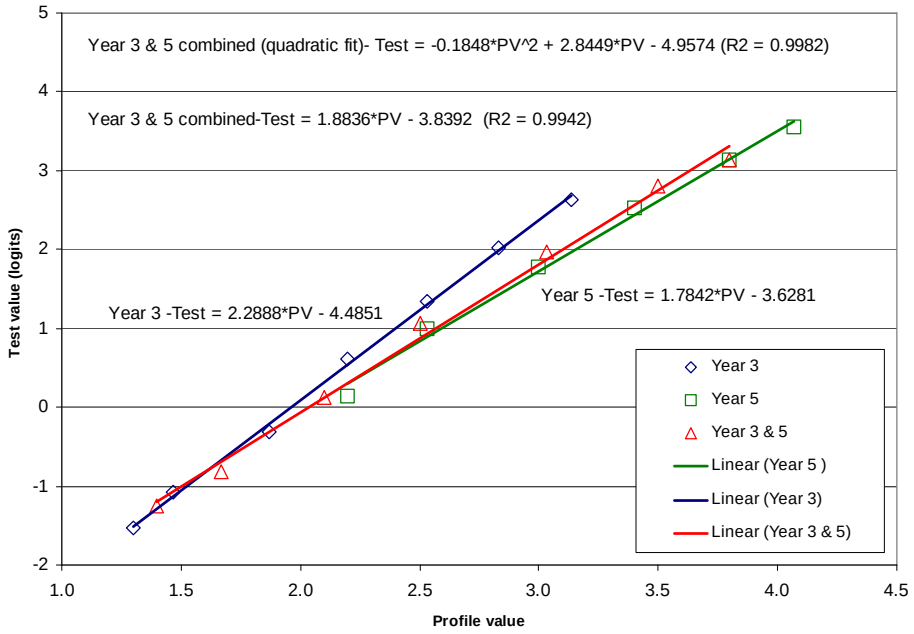
The next section of the chapter establishes the similarity of the equi-percentile equating result applied to the combined Years 3 and 5 to a Rasch model supported equating process. This comparison is made to justify the subsequent use of the Rasch model even though its application to the limited teacher-judgement data is problematic. That both equating processes produce similar results for the most part along the teacher judgement scale is offered as evidence that the conversion process of teacher judgement assessment scores to test scale values is robust.

Equi-percentile equating: Year 3 and 5 cases separately and combined.

Equi-percentile equating of the two scales, using the students common to both forms of assessment, is the simplest equating process to apply. Figure 8.1 illustrates the results of

equi-percentile equating for Years 3 and 5 separately. Using the seven percentile points (5, 10, 25, 50, 75, 90, 95), the Year 3 test scores are plotted against the mean profile level per student. These points turn out to have a clearly linear relationship. Ordinary least squares (OLS) regression is applied since R^2 is virtually 1.0, meaning that regressing Test on Teacher assessments produces the same result as regressing Teacher on Test. The Year 3 fitted line has a gradient of 2.28. Similarly the Year 5 percentile points are approximately linear with a gradient of 1.78.

Figure 8.1 Comparison of Equi-percentile equating by separate Year levels 3 and 5 with the combined data set for Years 3 and 5 - 1997 English.



The graphs suggest one of the potential contributors as to why the general relationship of the test scale to the profile scale is curved. Within a Year level the equating relationship is linear but the gradient of the relationship is diminishing with increasing Year level. It is a leap from the data of the two known Year levels to assume the likely gradients at other Year levels. At Year 2, however, it might be assumed to be steeper. For the intermediate Year 4 a gradient between those of Years 3 and 5 is logical. The same apparent variation of gradient with Year level applies independently in the mathematics data collected one year later.

It is known that as Year level increases, the span of the development range of students increases (Chapter 7) based on a teacher judgement assessment scale view. The SDs increase with Year level. The bulk of the class could be expected to be placed around the Year level mean, the judgement of which aggregated over many teachers is linearly increasing with Year level, as is the spread (see Figures 7.3, 7.9). As the spread increases, the learning-status-estimates further from the Year level mean are likely to be made with less detailed knowledge of the typical skill level of students at the extremes by teachers at that Year level. From

Figure 8.1 a Year 3 teacher sees a student with a test score of 2 logits (above the 90th percentile for Year 3) as at about 2.8 profile level units. The same test score position, now only above the 75th percentile for the Year 5 teacher, is seen as 3.2. The teacher assessment scales are Year level specific.

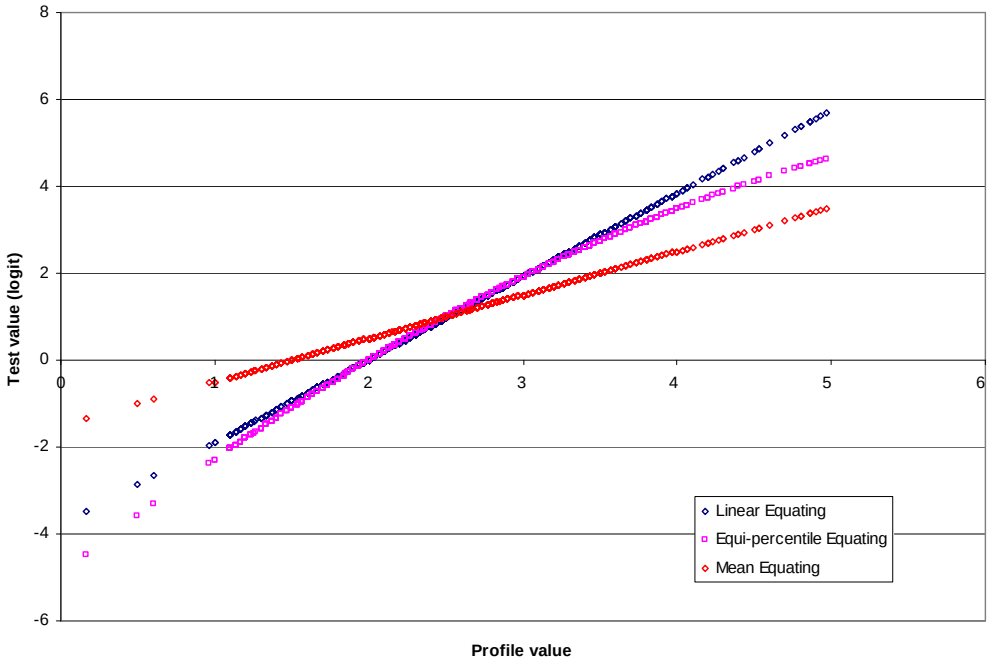
The purpose at this point is to estimate the general profile level to test score relationship over the profile level range from 1 up to 6 using the equi-percentile approach, even though the data for common students covers the range of 0 to 4 profile levels only. The common cases at Year 3 and 5 are well balanced on the spectrum of Year levels from 1 to 8 and thus profile levels from 0 to 6. The combined Year 3 and 5 teacher data sets (Table 8.3) have improved correlations with the test scores, relative to the two Year levels treated separately. The correlations are illustrated in Tables 8.2 and 8.3. For simplicity, the analysis continues on the basis of using the combined data for Years 3 and 5 to estimate the general equating relationship of profile level scores to test scores. An acknowledged consequence is a mild distortion of the equating relationship at each Year level. It will be shown later that the Rasch model equating produces a relationship for the two scales similar to that of the combined Year level equi-percentile method. As a result the choice of preferred process depends on the other benefits that might arise from the equating process chosen.

Using the combined set, based on equi-percentile points, the marginally best fitting equating relationship is curved. Figure 8.1 provides the linear and quadratic expressions for the lines of best fit. The curve itself is not shown to reduce visual complexity. The R^2s for the linear and quadratic fitted curves are virtually identical but the quadratic fit has a slightly higher R^2 value. Since the relationship of profile units to test score units is already known to be non-linear, the curvilinear relationship should be preferred for extension outside the Year 3 to 5 range. Ultimately a Rasch model equating is adopted as described below. It will be shown to be approximately identical to the equi-percentile equating but, prior to arguing the advantages of that process, three equating processes are considered to illustrate the ways in which those results differ. When it is shown empirically later that the non-linear equi-percentile approach approximates the Rasch model, the Rasch model equating result can be seen as proxied by the equi-percentile solution.

Comparing the equating results from mean, equi-percentile (linear) and equi-percentile (non linear) relationships.

Figure 8.2 illustrates the application of three equating processes to the 1997 data for Years 3 and 5 combined, based on plotting the result for each of the 1275 data points.

Figure 8.2 1997 Profile to Test scale equating by equating method, Year 3 and 5 data combined-English



As would be expected the mean equating process coincides with the other approaches only at the means of the two scales. The lack of equating of the spread generates an inadequate solution. The equi-percentile (linear) equating and equi-percentile (non-linear) equating produce approximately similar solutions over the range of 1.5 to 3.5 profile units. Outside this range the equating relationships spread apart. Of the three processes, only the non-linear equi-percentile equating is sensitive to the non-linear relationship between the teacher and test scales identified in earlier chapters.

Rasch model equating

An alternative equating process is applied, based on a Rasch model analysis of the full teacher assessed cases from Years 1 to 8 for 1997, 7871 cases altogether. In this process the three assessment strands in English are regarded as items. The item score values are obtained by deleting the decimal point. The items can take a value from 0 to 89, although the highest actual score is 70. The analysis is at the low limit of tolerance for a Rasch model using Winsteps and is not a conventional analysis. The approach uses the Rasch model akin to Wright's (2000) application of Winsteps to multiple regression (Bond & Fox, 2007, p. 203) and takes advantage of the capability of Winsteps to take 99 values for an item when a two-column format is used.

Three general options are available for the equating of the teacher assessments to the test scale under this process. One option is to use the 1275 common students (for the 1997 data) as person anchors for the full 7871 cases. The second option is to use a subset of the 1275

common points as anchors. The third option is to analyse the data set without anchors and then equate the teacher scale to the test scale using linear equating, effectively rescaling the teacher assessment logits to match the test logits.

Testing of all three options indicates that they produce approximately the same general equating result for the non-anchored points. However the first two options fix the student score on the teacher assessment to the test score (a logical expectation of the concept of anchor) for all or some of the points. In both options the anchoring pre-determines that the anchored cases will maintain their original relationships. This defeats the purpose of the investigation of the extent to which the two assessment process produce similar results since some (or all) cases have a predetermined relationship. For this reason, the equating exploration is developed using the third option, without anchors.

Non-anchored Rasch model equating

Appendix 12 contains the detailed statistical summaries of the fitting of the teacher data to the Rasch model for both the 1997 and 1998 collections. As indicated above the application of the Rasch model to such a messy data set is unconventional and produces a large number of poorly fitting cases. To complete the analysis using the Masters partial-credit model (Linacre (2000, p 300), a relatively large number of iterations were required (741 for 1997, 275 for 1998). In this less conventional approach there are 90 categories of partial credit for three items in the English case and for four items for the mathematics case.

In Appendix 12, Table A12.1 the mean square infit value for the three items for 1997 is 0.94. The actual infit values for the three items are 0.72, 0.91, and 1.18. In general terms these items fit the model (between 0.7 and 1.3) and have an item reliability of 0.92. The limited number of items and the person infit mean-square mean well below 1.0 indicates a high degree of over-fitting cases. Boxplots of the distribution of person infit mean-squares are shown in Appendix 12, Figure A12.1. These illustrate the high degree of skew towards 0 with the median well below 1 and the wide spread of values. As Year level increases the spread increases. Inspection of the cases at each end of the spectrum offers some understanding of the reasons. Those cases with infit values at or near zero have no between-strand variation, a consequence of teachers seeing the progress within each strand as equivalent. Cases at the upper end (higher infit values) have one strand where the value varies by 0.5 profile units or more from the other two. The Year level trend reflects the increasing variability between strand assessments as Year level increases, that is the teachers discriminate more between developmental status in each strand as the Year level of the student increases.

Table A12.3 shows high negative residual correlations between items, further evidence of over-fit and the variance in the data explained by the model measure is very high at 97.7% (Table A12.4) supporting the purported unidimensionality of the data. There is lack of randomness in the data due to the few items and commonly high correlations in the ratings for each person on each item due to their being at roughly the same point of development on each strand. Based on Linacre (1999) "some randomness is needed in the data in order to construct a measurement system…In the case of local independence, however, the fit interpretation is reversed. The closer the data comes to the perfect Guttman pattern the less local independence there is, and so the worse the fit." (Linacre, 1999, p. 710) In this application of the Rasch model the purpose is to approximate fit to the model sufficiently to bring the teacher assessment data into an arrangement that assists the transformation of scores from the teacher scale to the test scale. Given the very few items, can the use of the model be justified?

It will be shown below that the relationship estimated for the cases coincides for large segments of the teacher scale with the equi-percentile (non linear) solution. On this basis the measure values estimated for items and persons are proposed as having sufficient heuristic value to explore the modelled data further. One reason is the benefit of the Rasch model in estimating measurement error for each case.

The comparable set of Rasch model fit statistics for the 1998 (mathematics) data are provided in Table A12.5. More items are provided (4 as against 3) with an infit mean square mean value of 0.75 but a narrower range of infit values. Item reliability is reported as 1.00 with a much higher separation statistic of 24.57 (compared to 3.47 for English). The boxplots in Figure A12.2 show a high degree of overfitting, consistent with the lower infit mean square mean relative to the English data. The SD of the infit is less; the reduced spread is observed in the boxplots, as is the trend of increasing spread with Year level. High negative residual correlations are found (Table A12.7) but with lower values than for English. Table A12.8 indicates support for the premise of unidimensionality but with low local independence as for English, due to the inter-related nature and limited number of items.

Item/Strand difficulties

The strand difficulties for the two teacher judgement collections are shown in Figure A12.3. While the difficulties are presented side by side the scales should not be assumed to be the same. The figure illustrates the closeness of the difficulties of the three items for English and the greater spread of the difficulties of the mathematics strands. The appendix provides more comparisons of strand issues that are not dealt with here. Based on the test-teacher correlations reported earlier (English-Literacy 0.66, Mathematics-Numeracy 0.57- both highly significant at the 1% level), continuing the analysis using the total test score and profile average over strands has reasonable face validity even though the teasing out of strand detail might be problematic.

Converting the Unanchored Teacher Rasch measures to the test scale

The Rasch measures obtained in Appendix 12 are converted to the test scale by the following process. Using the 1997 data, the teacher-assessed cases common to the tested population (at Years 3 and 5) are selected. The mean of the teacher assessed score for this group has a value of –1.64 logits and an SD of 1.33 compared to the test assessed scores that have a mean of 1.06 logits and a SD of 1.37. These are found in Table A12.2. The teacher measures from the Rasch analysis are converted to the test framework by making the mean and SD of the common cases equivalent to the test mean and SD by the standard procedure (teacher means rescaled to the test mean and spread in proportion to the ratio of the SDs to make the means and SDs of both data sets identical). The result is confirmed in the third column of Table A12.2. The full teacher assessed set, that is the additional students at other Year levels, are re-scaled on the same basis to create a set of 7871 cases with a mean of 1.23 logits and a SD of 2.07 logits. This compares with the original values of –1.47 and 2.00 in the second column.

This process has re-scaled the length of the teacher assessment scale logit. Because it is used as key part of the comparison of the individual common cases, the error of measurement is also rescaled to the test logit scale. In the case of the 1997 data this re-scaling of the error makes little difference to its values (see the right column). However the same process applied to the 1998 data (see Table A12.6) produces an increase in the error of measurement (from a mean of 0.18 for the common cases to 0.27 when rescaled) due to the differences in SD (1.44 test, 0.99 teacher).

The results of the rescaling are shown in Figures 8.3 and 8.4. The original teacher Rasch measures in logits on the vertical scale are summarised at their profile level values. This is the lowest line. The re-scaling lifts the line to a new position and rotates it slightly and represents equated test logits on the vertical scale. In the same figure the individual Year 3

and Year 5 equi-percentile equating lines are shown as lines of indicated gradient. In addition the equi-percentile equating line using the fitted quadratic curve for the combined Year 3 and 5 data is shown. This line is approximately identical to the Rasch model for score conversion for large portions of the teacher profile scale. The two curves deviate below 1.3 profile units and above 4.0 units. The Year 3 and 5 lines appear to touch (or follow) the coalesced curves as approximate tangents. This indicates that the general conversion (whether Rasch or equi-percentile) over the full profile level range is sensitive to the changes in scale conversion values as the Year levels of teachers increase.

Figure 8.3 1997 English Teacher assessments – Conversion of Profile to Test Scale result: Independent Rasch model.

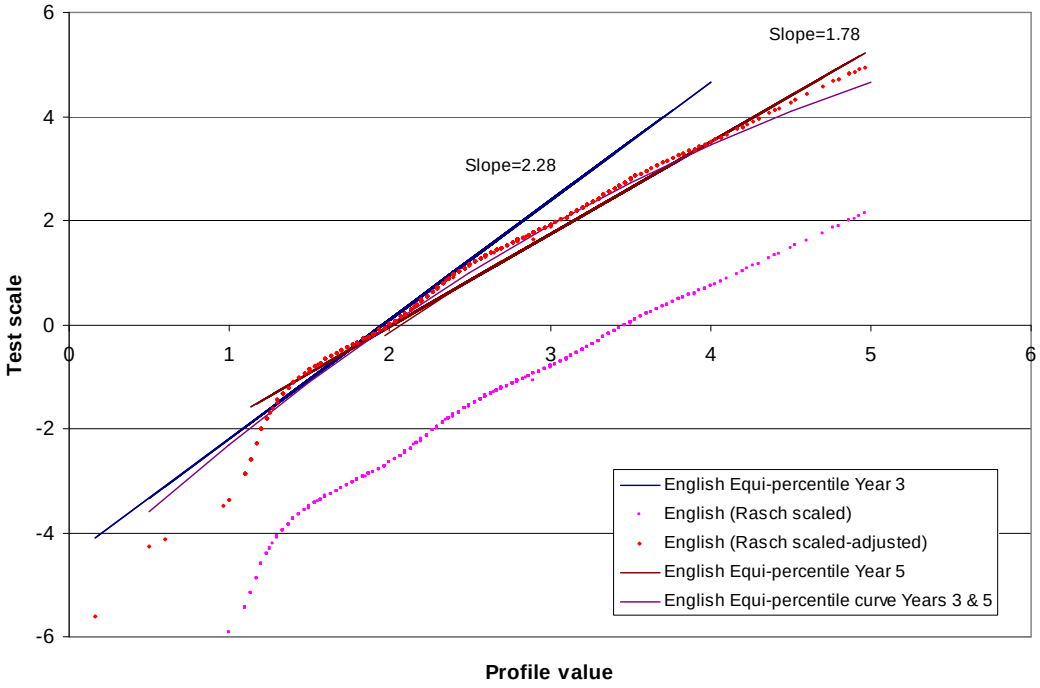
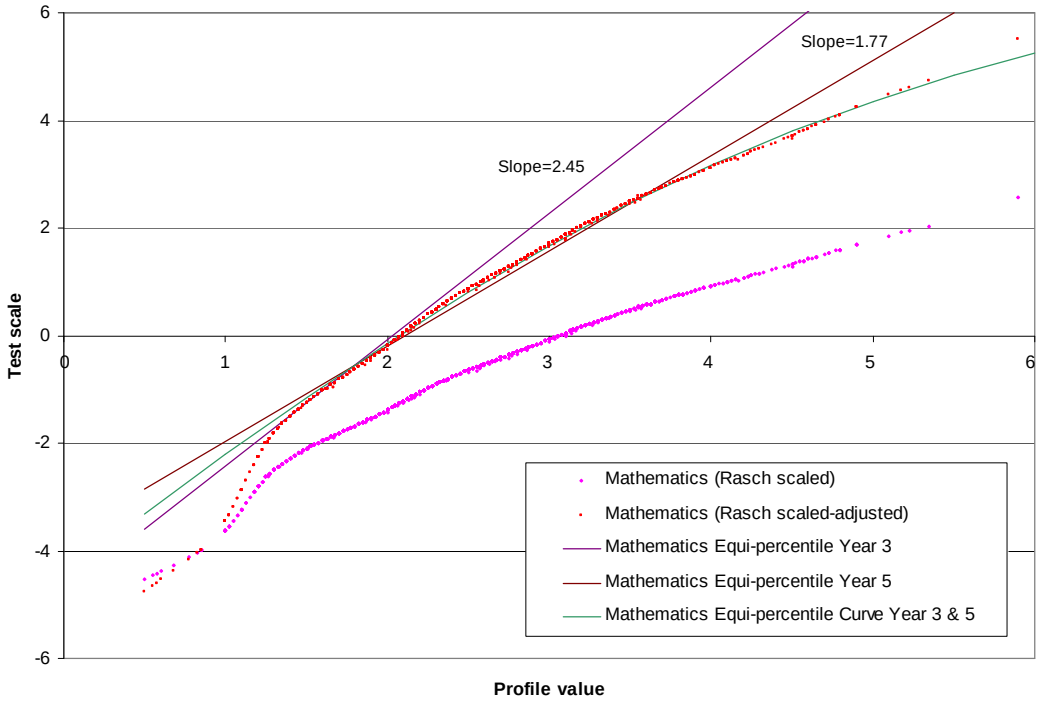


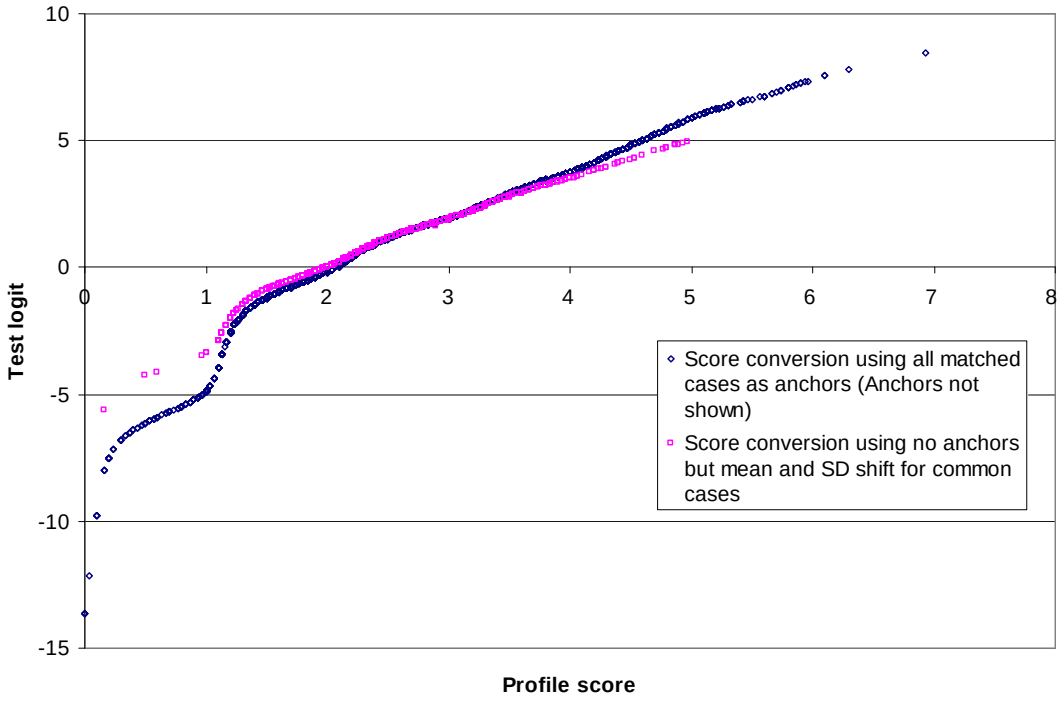
Figure 8.4 1998 Mathematics Teacher assessments – Conversion of Profile to Test Scale result: Independent Rasch model



As indicated earlier this general solution (whether Rasch or equi-percentile) averages the conversion over the profile level range, relative to the conversion relationship at each year level (the linear alternatives). As the investigation is concerned with the broad relationship only, and as the conversion can be applied generally over the full Year level range 1 to 8, the Rasch model profile level to logit conversion is adopted. That it is approximately identical to the combined Year 3 and 5 equi-percentile curve supports the adequacy of the Rasch conversion, even though the model application is a little unconventional with many over-fitting cases

Finally, in completing the initial re-scaling of teacher assessments to the scale of the test, the adequacy of the re-scaling of the teacher assessments through the Rasch model linear equating is confirmed by comparison with independently applied anchored equating. Figure 8.5 compares the linearly re-scaled result with the result of using all the common cases as anchors but suppressing the plotting of the common case points. Since the logit values obtained for the anchored approach are already directly linked to the test scale values no additional re-scaling is required. The conversion lines are very similar, as above, in the range 1.0 to 4.0 on the profile level scale.

Figure 8.5 A comparison of the final result of the unanchored conversion of the teacher scale to the test scale compared to the anchored result



Comparing Teacher and Test Assessments for Common Students with Teacher Assessments Re-scaled.

1997

The 1275 cases where students have both a test assessment and a teacher assessment, now converted to the test scale, can be compared. Each student with a test and teacher assessment in test logits also has error of measurement estimates for both assessments. Using the approach described by Wright and Stone (1999) for items and Bond and Fox (2007) for persons, the control lines for 95% confidence for each assessment on the two scales can be applied to the scatter plot of data points shown in Figure 8.6 in order to examine the invariance of the person measures across assessment types.

For a 95% confidence range, control lines are set so that the perpendicular distance of the control line from the 45 degree identity line is 2T, where T is the unit of error related to the difference between the two measures for the case. The standard unit of error of the difference calculated on either axis for person$_i$ is

$$S_{12i} = (s_{1i}^2 + s_{2i}^2)^{1/2}$$

with $T_{12i} = \left[(s_{1i}^2 + s_{2i}^2) / 2 \right]^{1/2} = S_{12i} / \sqrt{2}$.

The upper control line can be plotted with the coordinates

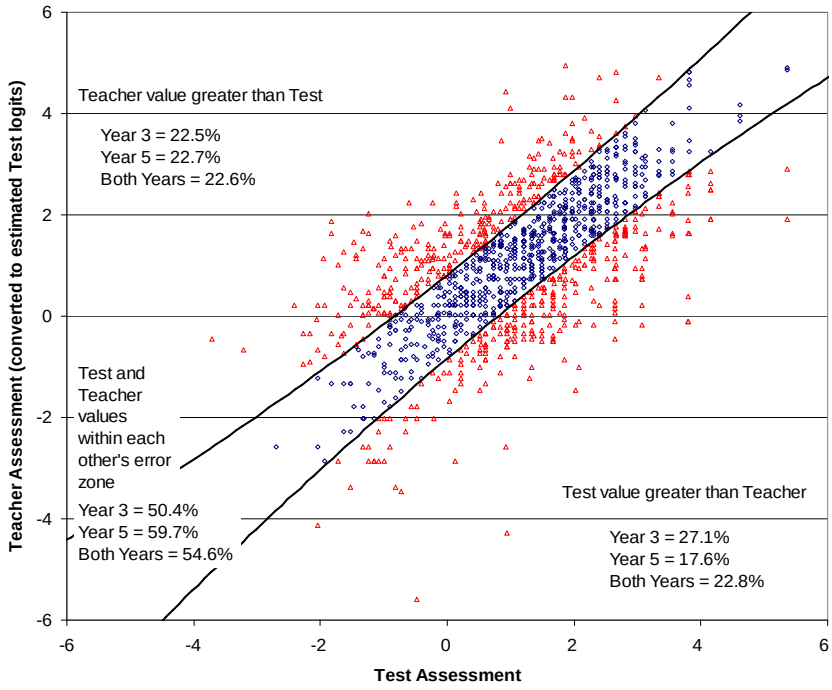
$$X = d - 2S_{12}/2 = (d_1 + d_2 - 2S_{12})/2$$
$$Y = d + 2S_{12}/2 = (d_1 + d_2 - 2S_{12})/2$$

where $d = (d_1 + d_2)/2$, d_1 and d_2 being the values of the measures on each axis, and S_1 and S_2 being the standard errors of measurement. (Based on Wright & Stone, 1999, p. 71).

The lower control line is symmetrical and plotted by reversing the X and Y coordinates.

The resultant (idealised) control lines are shown in Figure 8.6. The number of cases that can be regarded as matched within a confidence range of 95% of the errors of measurement, fall within the boundaries of the control lines, using the rescaled error of measurement values for the teacher measure. The estimates of the proportions of cases that can be considered as equivalent on both measures for English/Literacy are shown in Table 8.4.

Figure 8.6 1997 English/Literacy - Scatterplot of Teacher assessment and Test assessment invariance



The overall match rate of the two assessment processes of the combined Year 3 and Year 5 data is 54.6% of the cases. The proportions of cases above the upper control line and below the lower control line are equivalent, indicating that it is as common overall for the teacher score to be above the test score as it is for the test score to be above the teacher score when the scores do not match.

Table 8.4 1997 Comparison of Teacher and Test assessments of common students

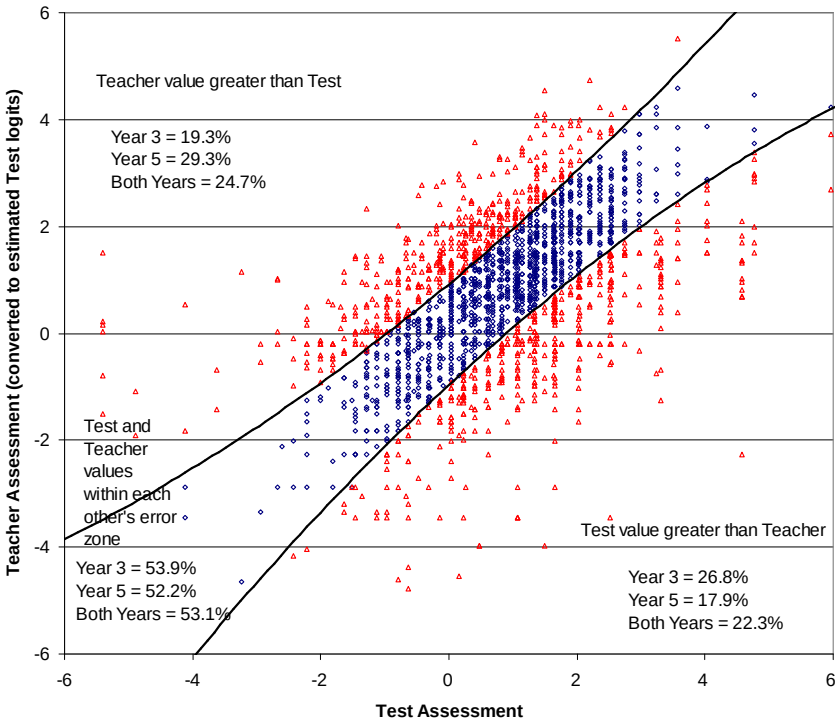
	Year 3	Year 5	Both Years combined
Teacher assesses student above Test	22.5%	22.7%	22.6%
Test assess student above Teacher	27.1%	17.6%	22.8%
Both processes within error zone of each other	50.4%	59.7%	54.6%

When the cases are examined for the specific Year levels the match rates vary. More Year 5 assessments match than do Year 3 assessments. (Part of the reason for this relates to the greater closeness of the original Year 5 only conversion scale to the combined conversion scale.) The rate of teachers assessing significantly above the test score is constant by Year level at 23% rounded. At Year 3 the test provides a score higher than does the teacher in 27% of cases. At Year 5 this test above teacher score rate is a lower 18%, with an increased match rate to 60%. The other possible sources of assessment error in the test process and the teacher assessment process, beyond the measurement error estimated within the Rasch model, are considered later.

1998

The situation for the 1998 data is summarised in Figure 8.7 and Table 8.5. The rescaling of the teacher assessments to the scale of the test required the size of the teacher assessment logit to be increase to a greater extent than for the 1997 data. As a consequence the teacher error estimates reflect proportionately adjusted values to match the test scale logit. The scatter plot patterns for Mathematics/Numeracy are similar to those for English /Literacy. The proportion of cases that fall within the control lines is 53%, slightly less than for English/Literacy. As for English/Literacy the cases outside the control lines are balanced at about 22-24%. For the individual Year levels the cases within the control lines remain about at 52-54%.

Figure 8.7 1998 Mathematics/Numeracy - Scatterplot of Teacher Assessment and Test assessment invariance



In mild contrast to the English/Literacy situation the proportions of teachers assessing the student more highly than did the test are greater for Year 5 than for Year 3. As a corollary, the proportion of test scores higher than the teacher is greater for Year 3. The equating of teacher and test scores using the Rasch model approach has averaged out the steeper conversion line of teacher assessments to the test scale that applies to Year 3 in isolation. If the steeper Year 3 conversion line were used it should make no difference to these proportions, as it can be seen from Figure 8.4 the conversion line follows the Year 3 linear gradient for most of the range in which Year 3 assessments are placed.

Table 8.5 1998 Comparison of Teacher and Test assessments of common students

	Year 3	Year 5	Both Years combined
Teacher assesses student above Test	19.3%	29.3%	24.7%
Test assess student above Teacher	26.8%	17.9%	22.3%
Both processes within error zone of each other	53.9%	52.2%	53.1%

Summary of rates of teacher assessments matching test assessments

Accepting the assumptions and results of the Rasch model equating process, teachers' assessments match test assessments in just over 50% of the cases, allowing for errors of measurement. This degree of matching occurs in two independent sets of assessments one year apart. By categorising the scale on both the test and teacher axes into 1 logit categories,

a Cohen's Kappa value of just above 0.4 is obtained. This is regarded as a fair to moderate agreement only (Altman, 1991). The combining of well-calibrated, moderately-calibrated and poorly-calibrated teachers into the one analysis leads to this relatively low agreement rate. Later in the chapter it is shown that at some school sites much higher agreement rates apply.

The spread of the error zone reinforces that all assessments are made with error. One element, the modelled error, is used to set the control lines. The modelled error, the estimate of the range within which the actual score might lie when the assessment process fits the Rasch model, is quite large. The general size of mean measurement error for the tests in 1997 is between 0.33 to 0.37 test logits (Table 6.2) and 0.27 for the teacher assessments (Table A11.2). The 1998 equivalents are 0.39 to 0.49 for the tests (Table 6.3) and 0.28 for the teacher assessment (Table A11.6). An error of 0.3 test logits is equivalent to about 7 to 9 months of learning development based on Hungi (2003), and is a direct consequence of the relatively small number of items routinely used in testing situations.

Part of the complexity in making the comparison of teacher and test assessments is the use of a single standard procedure, assumed to work consistently in all applications (the test) with a looser but still standardised process open to more varied application and interpretation (the teachers). Assuming that the processes are assessing the same learning trait, more replications of the assessments for individual teachers would provide the data to tease out the potential sources and causes of disagreement in the assessments. The current data sets cannot offer much insight into the likely reasons when assessments do not match. However, within the error tolerances of the assessment processes and the model for equating, slightly more than 50% of the students can be regarded as having invariant assessments across forms. As will be revealed later, there is evidence that at some individual school sites (as proxies for teachers) the number of cases that match is low but the correlation between the two assessment processes is high. This indicates that considering the matching alone is an inadequate basis for comparison.

There are very few published examples of the relative performance of teachers and tests in assessing students using common scales. Examples in Chapter 4, in particular Tables 4.2 to 4.4 and Figure 4.3, indicate degrees of match but in most cases using a much less precise scale The broader the scale unit, the greater the chances of teacher and test assessments matching. An estimated match rate at Key Stage 3 of 61.5% for English and 70% for mathematics (Figure 4.3) uses the very broad unit of one Key Stage level. Teachers in England were also assessing to a more explicit framework; which was also used by test designers to develop the tests. One of the sources of variability is the teacher. Can the data provide an insight into the effect of teacher assessment skills on the extent of match?

Estimates of between teacher differences in matching

As indicated earlier there is no way that individual teachers in the data set can be identified. By design the students assessed by particular teachers cannot be grouped together. However as a consequence of the collection process it is possible to aggregate data at each school site, after recoding site codes to remove their specific identities. Teachers provided, on average, 5 student assessments each. For sites with more than 5 students per Year level, the site data are for multiple teachers. Thus the teacher-assessment test relationships for small groups of teachers at each site can be established. The general match rates in the previous sections can then be re-examined within and across sites.

Figures 8.8 and 8.9 show the distributions of the match rates (i.e., invariance within measurement error) based on sites converted to the estimated numbers of teachers at each site. The mean match rate at the site is ascribed to all teachers. In practice teachers at a site would also be likely to vary in their match rates.

Figure 8.8 Match rates 1997 - English/Literacy

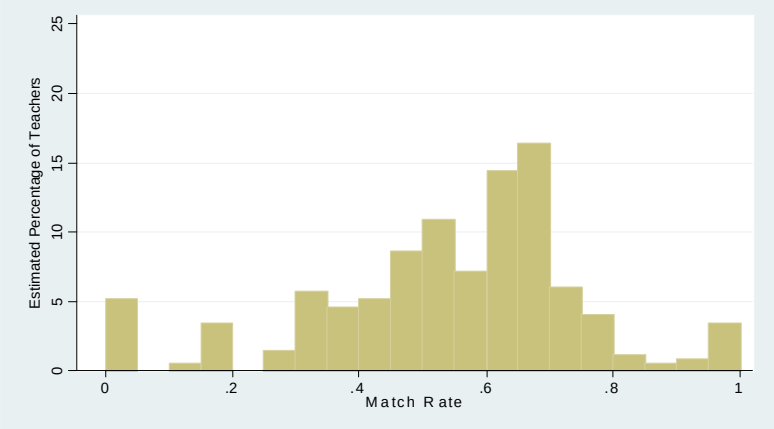
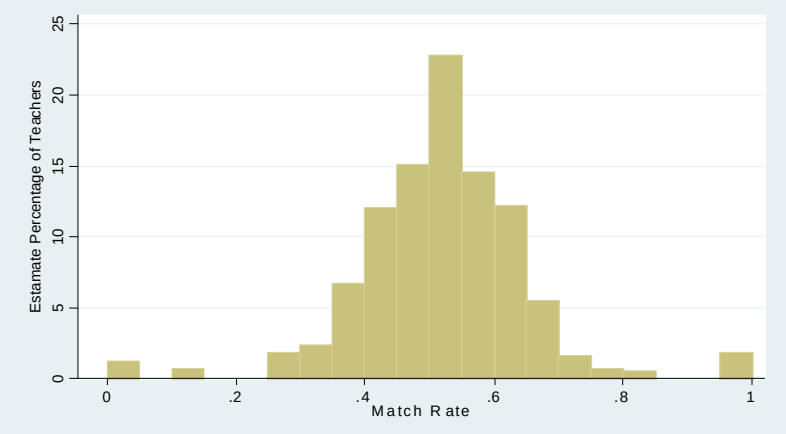


Figure 8.9 Match rates 1998 - Mathematics/Numeracy



The distributions appear to be different. This may be learning area related, although as raised earlier there is the possibility of non-random loss of data as a result of the process to connect

test and teacher judgement assessment for individual students. The English/Literacy distribution is less concentrated around the mean match rates, with proportionately more cases with no match and with high match. Although both learning areas have similar overall match rates, the mode values differ (near 0.7 for English/Literacy, nearer 0.5 for Mathematics/Numeracy). It appears that there is greater variability in the match rate of assessments in English/Literacy than in Mathematics/Numeracy. Some English/Literacy teachers are well aligned to the test scale (above say a match of 0.7). Fewer Mathematics/Literacy observations were as well matched. However some English/Literacy assessments are very poorly aligned (below a match rate of 0.3) relative to Mathematics/Numeracy.

Six case studies for English in Figure 8.10 and Table 8.6 illustrate an approximation of what the situation might look like when multiple students for individual teachers are examined. The cases are selected on the basis of relatively high numbers of students at a site (above 13 implying at least three teachers) and for a range of match rates. Match rates are calculated as the proportion of measurably invariant assessments relative to the total number of assessments. The highest match rate of the selected cases (site 1687) is 0.92, the lowest 0.16 (site 2777). The scatter plots all have positive slopes and positive correlation coefficients. For site 1222 there are some outlying cases. Sites 1936 and 2777 have quite varied matching rates (0.7 and 0.16) but high correlation coefficients (0.95 and 0.79). These cases highlight the matters raised in Chapter 4 on forms of matching. Intercepts and gradients using TLS/Deming regression advocated in Chapter 4 (as distinct from OLS) are included as broad indicators of the variation in teacher assessment matches with the test score across sites.

Figure 8.10 1997 English/Literacy: Comparison of Teacher assessments (Rasch model equated) and Test Model assessments at selected sites.

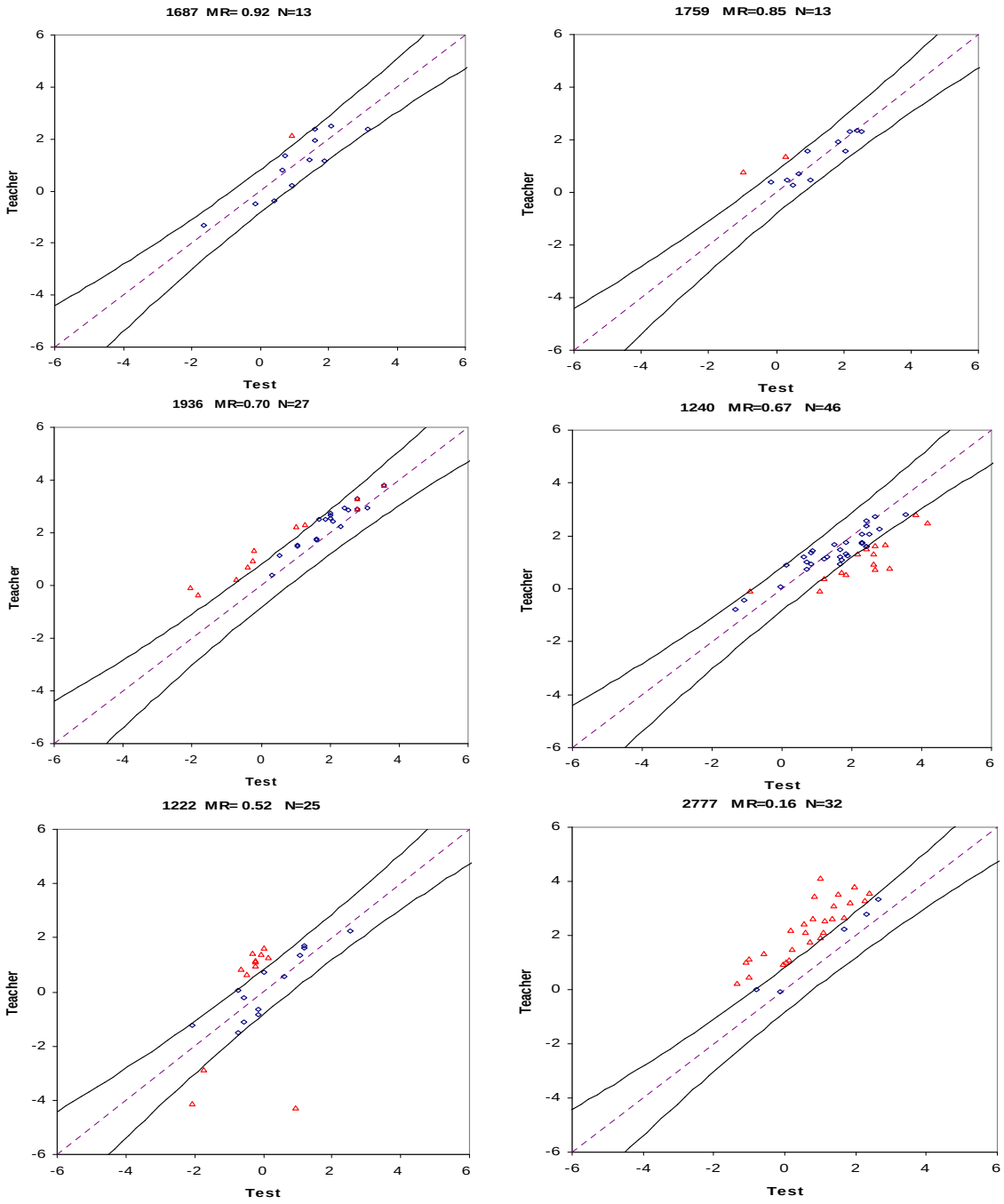


Table 8.6 1997 "Deming" Regression and Kappa values for Teacher and Test assessments of common students at selected sites

Site	Test Mean	Teacher Mean	Sample size	Corre-lation	Match Rate	Variance ratio	Intercept	Std. Error	Slope	Std. Error	Kappa
1687	1.05	1.07	13	0.85	0.92	0.96	-0.07	0.34	1.08	0.27	0.45
1759	1.04	1.28	13	0.78	0.85	0.67	0.59	0.32	0.65	0.19	0.50
1936	1.27	1.92	27	0.95	0.70	0.44	1.00	0.11	0.73	0.05	0.58
1240	1.75	1.30	46	0.84	0.67	1.82	0.11	0.13	0.68	0.06	0.54
1222	-0.13	0.07	25	0.54	0.52	2.87	0.44	0.56	2.93	1.47	0.41
2777	0.73	2.13	32	0.79	0.16	0.12	1.47	0.13	0.91	0.09	<0

The assessments in the lower panels have fewer cases within the control lines but have medium to high correlation coefficients. Deming regression analysis (allowing error on both axes) indicates some reasons for mismatch related to teacher calibration to the test scale.

At site 2777, with a low match rate of 0.16, many cases are above the upper control line and thus do not meet the criteria for a high match of assessments. The correlation is high, the slope for the regression close to 1 (0.91 with low standard error, 0.09) and with an intercept on the teacher axis at 1.46 (standard error of 0.13). Taken together these data points imply a high degree of calibration to the test scale, but with teacher assessments consistently of the order of 1.5 logits above the test. From the slope (0.91) it is seen that the scale range for teachers is slightly narrower than the test scale. The teachers are however clearly following the scale of the test but their assessments are displaced consistently above it.

Kappa values are obtained by categorising the assessments into 1 logit wide categories on each scale. Apart from site 2777, positive values above 0.4 are obtained, indicative of a fair to moderate agreement (Altman, 1991). For site 2777, the Kappa value of less than 0 implies a lower than chance match. If however the scale categories for the Kappa calculation are re-categorised after a 1.5 logit shift down on the teacher scale, the Kappa value becomes 0.58, equivalent to the highest Kappa in Table 8.6. A major reason for the assessments not matching at this site is the teachers systematically assigning higher values to students relative to the test assessments, leading, to an over-estimation of scale positions by teachers relative to the test. Rescaling of all cases down by 1.5 logits leads to a match in most cases, implying these teachers are calibrated to the scale but systematically over estimate scale positions.

Most of the case study sites show consistency in assessments within a school, even though the test and teacher assessments may not be measurably invariant. The correlation coefficients are at or above 0.78, except for site 1222. This puts the selected case studies mostly above the overall correlation coefficient for the full 1275 cases of 0.66. This reinforces that these are selected sites (on the basis of the varying match rate across the matching scale and relatively high correlation coefficients) and thus do not necessarily represent the general pattern. However the scatter plots suggest it is possible to have multiple teachers at a site (*n* estimated to be 6 for site 2777) assess consistently at Years 3 and 5. That is they all seem to follow the same general understanding of learning status even though this common understanding is systematically displaced from the test calibration. This observation can be made of all exemplar sites except 1222 where more outliers indicate greater variation in teacher and test perspectives.

In the case of site 1222 the Deming regression indicates through the high slope value (2.92) that teachers have a markedly wider scale range for their assessments than does the test. The

reverse applies for site 1240; a narrower teacher than test range. Both examples illustrate that the order of students can be approximately consistent for the teachers and the tests but without a calibration process to ensure that the scales are seen to have equivalent units, the usefulness of the teacher chosen scale value, as a description of learning status, is diminished. Both examples, when seen relative to the other examples, offer hope that it is feasible to attempt to train teachers to locate their perceptions of learning development on a common scale.

Table 8.6 provides the Kappa value for strength of agreement between the two assessment methods at each site. Assessments are categorised into categories 1 logit wide on each axis to calculate Kappa. Based on Altman (1991) most agreements are either moderate (0.41-0.60) or fair (0.21-0.4). For case 2777 as explained above, adjusting each teacher assessment downwards by 1.5 logits leads to a revised Kappa of 0.58, confirming that the main reason for assessments mismatching is the systematic misalignment of teachers to the test scale. When a statistical adjustment is made the match rate becomes 0.94, only two cases remain outside the confidence limits. The case studies illustrate that for a number of teachers in English/Literacy the test and teacher scales are closely related even when the match rate is low. At some sites the scales correlate less well. The potential is there however to study those teachers who align well, to attempt to understand and develop processes to train other teachers to be so aligned to the test scale.

Figure 8.11 1998 Mathematics/Numeracy: Comparison of Teacher assessments (Rasch model equated) and Test Model assessments at selected sites

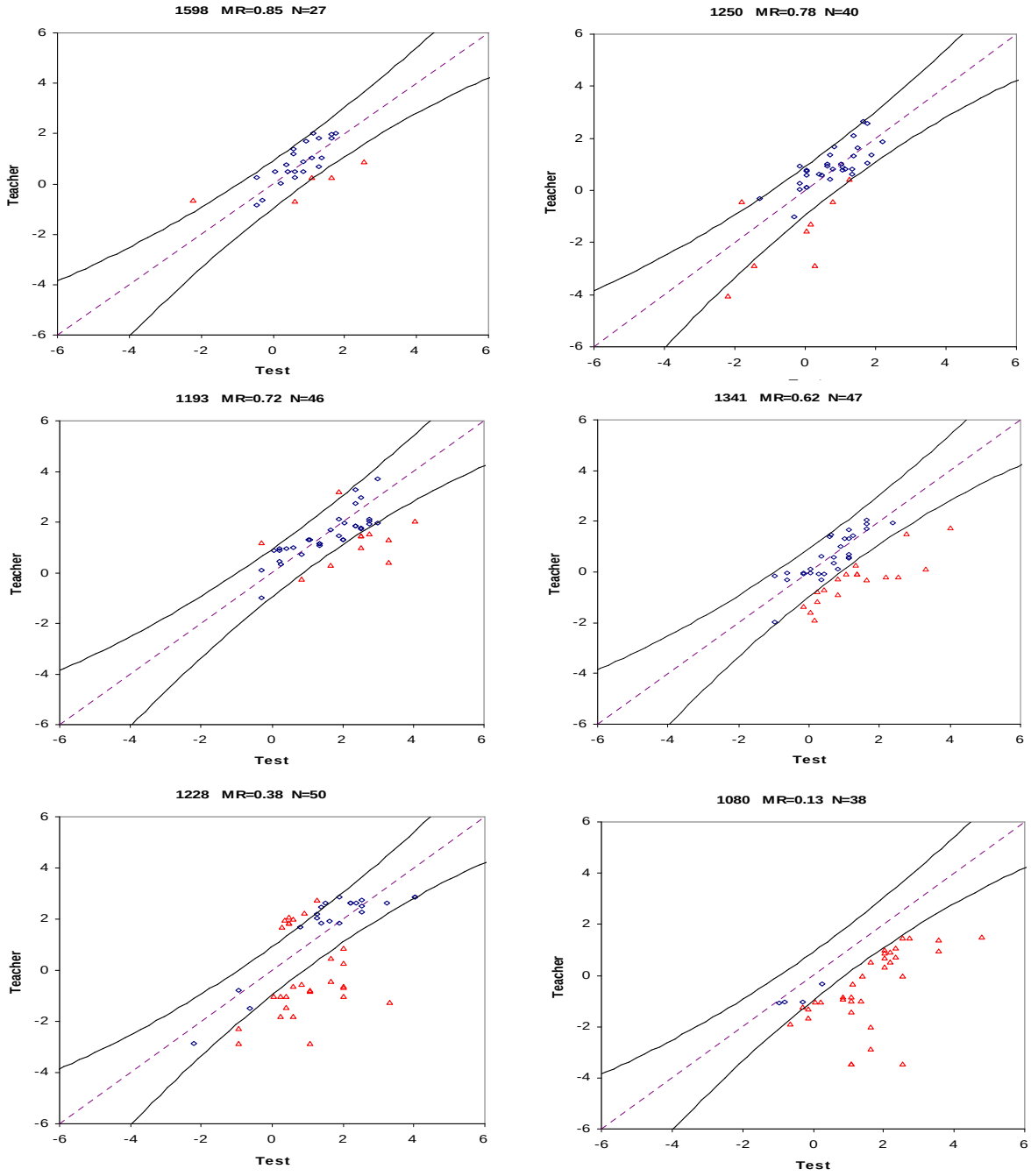


Table 8.7 1998 "Deming" Regression and Kappa values for Teacher and Test assessments of common students at selected sites

Site	Test Mean	Teacher Mean	Sample size	Correlation	Match Rate	Variance ratio	Intercept	Std. Error	Slope	Std. Error	Kappa
1598	0.72	0.72	27	0.65	0.85	1.00	0.08	0.37	0.89	0.39	0.41
1250	0.52	0.40	40	0.75	0.78	1.66	-0.46	0.28	1.66	0.26	0.42
1193	1.81	1.47	46	0.68	0.72	1.53	0.12	0.27	0.74	0.16	0.36
1341	0.89	0.25	47	0.54	0.62	12.71	-1.30	0.37	1.74	0.34	0.37
1228	1.22	0.66	50	0.57	0.38	3.41	-2.25	0.85	2.38	0.58	<0
1080	1.34	-0.52	38	0.56	0.13	7.07	-3.02	0.97	1.80	0.60	<0

Figure 8.11 and Table 8.7 show site case studies for the Mathematics/Numeracy assessments. Site 1598 represents 27 students assessed by an estimated 5 teachers. All but 4 assessment results are invariant. The Deming regression slope is close to 1 (0.89) and the intercept very close to 0 (0.08). These teachers could be regarded as approximately calibrated to the test scale. The correlation (0.65) indicates that further training might be required to improve the linking of teachers to the test scale along with examination of aberrant cases to clarify which assessment process is the less accurate.

As for English/Literacy, a low match rate does not imply a low correlation. Site 1080 with 38 cases (estimated 7 teachers) has a low match rate of 0.13 but a correlation coefficient of 0.56. The assessments are generally below the identity line, with most cases below the lower control line. Some of the outliers would suggest a poor relationship to the test scale. However ignoring the worst four outliers (particularly if they were to belong to just one teacher) produces a scatter that indicates a site consistency at least. All teachers at this site have somehow developed a consistent view among themselves of the use of the teacher assessment scale, and thus all appear to under-estimate their students learning development when the test scale is adopted as the standard.

The absence of identified individual teacher cases (and too few assessments per teacher even if they were identified) means that the judgement consistency of individual teachers cannot be observed. Analysis at a site level provides a deeper appreciation of the possibilities for teachers to become calibrated to the scale of appropriate test measures on common dimensions of learning. The cases studies are not necessarily representative of all sites but illustrate that much deeper understandings of the assessment behaviour of teachers are obtained when a site view is taken. An individual teacher view should be even more informative. Even though a site may have few assessment cases that are invariant (within error), there are sites where the teachers appear to be assessing consistently and bear a common - but displaced relationship - to the test scale. Such behaviour if confirmed elsewhere would provide a basis for building a common scale approach to student developmental assessment where teacher and test assessments could be constructively blended to provided an integrated approach to classroom assessment.

The conversion of teacher to test scales in this study is normative. It assumes the mean practice of the teachers indicates where the test and teacher scales should equate. An alternative analysis focused on specific skills and behaviours might establish a more appropriate relationship of test scores to a teacher scale or vice versa. Alternative processes to set the linkages (Hattie & Brown, 2003) might then provide a criterion basis for linking the scales. Such equating would then enable studies to establish (say in Victoria) a better

indication of the extent to which individual teachers are directly calibrated to, or consistently displaced from, the test scale allowing then the potential for individual teacher re-alignment.

Having established a scale linking process through the common cases at Years 3 and 5 it is possible to explore (speculatively) the assessments in the full range from Year 1 to 8. The next section addresses this more global comparison.

Extending the comparison of Teacher and Test assessments beyond Years 3 to 5

Comparing Teacher assessments to the Years 1 to 8 Test data model.

Using the common cases as a basis, the full set of teacher assessments is converted to estimated test logits. This conversion is made by applying the rule established in the common cases to the full range of teacher judgement assessments. The teacher data are thus expressed in test logit values rather than in profile level units for each student. These re-scaled teacher assessments can be compared with the test model developed in Chapter 6, an estimate of what the test data might look like, based on the best estimates of the trajectories of growth with age/Year level.

Figure 8.12 1997 English/Literacy-Mean of Teacher assessments (Rasch model equated) and Test Model assessments compared, by gender and Year level

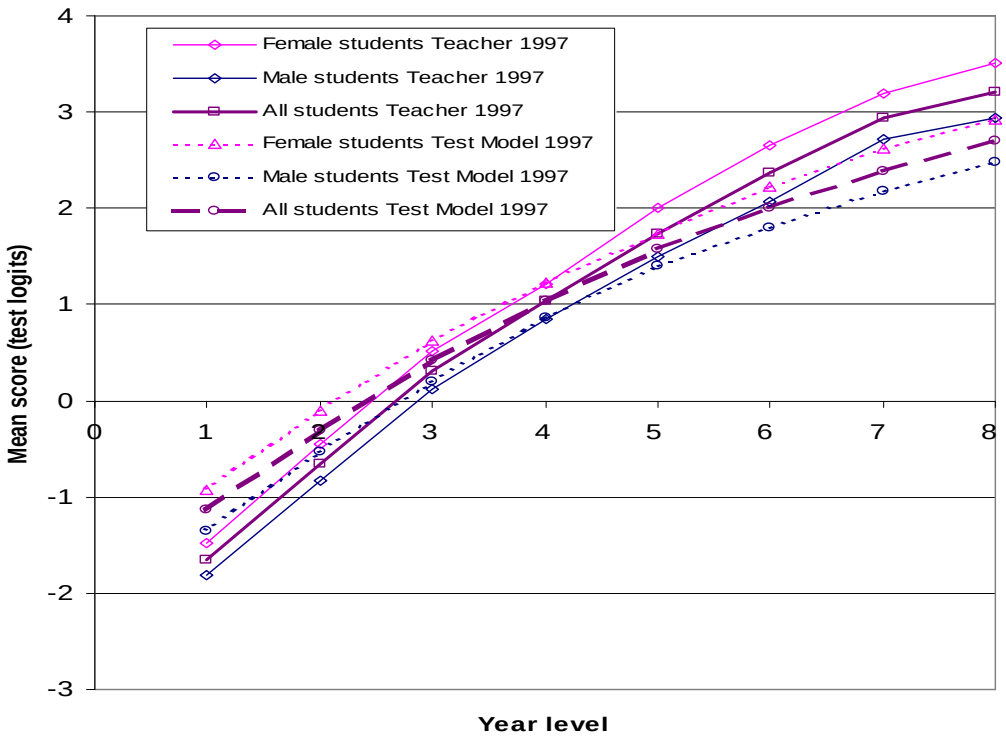


Figure 8.12 compares the two assessment processes for English/Literacy by Year level and gender. It was established in Chapters 6 and 7 that there were morphological similarities in summaries by Year level and gender for teacher assessments of English when compared with

the Literacy test. At Year 4, the notional average of Year 3 and 5, and thus at the point where the teacher-assessed cases are equated to the test scale, all students, male students and female students coincide for both assessment processes. Thus while the equating is performed on an 'all-students' basis, the gender patterns from both assessment processes are very similar (and quite different from the pattern for mathematics shown later). This provides evidence that the teacher assessments describe the Year 4 population in the same way by gender as do tests.The relative gender relationships apply from Year 1 to Year 8. The trajectories of the teacher assessments and the model test data however differ. It is not easy to establish the extent to which the trajectory difference is an artefact of the multiple assumptions that led to the establishments of the test model (only Years 3, 5 and 7 are actual data) and/or the process applied to convert the teacher assessments to the test scale. A later section will compare the data sets where the trajectories are also equated.

Figure 8.13 1998 Mathematics/Numeracy-Mean of Teacher assessments (Rasch model equated) and Test Model assessments compared, by gender and Year level

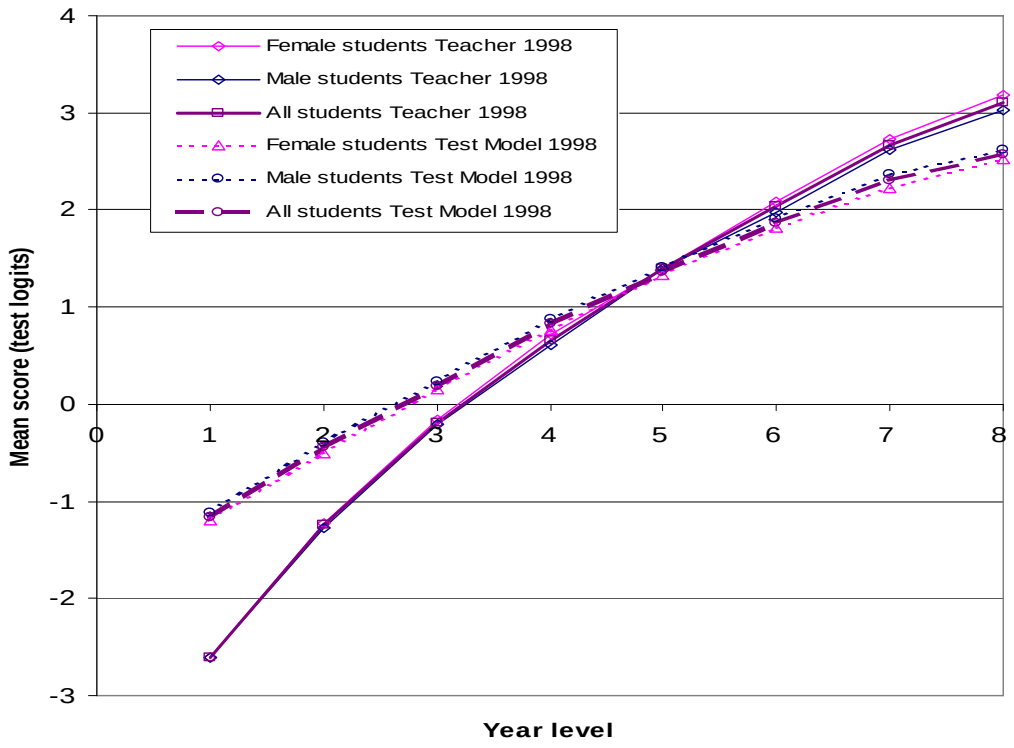


Figure 8.13 presents the Mathematics/Numeracy comparison over 8 Year levels. The general trajectory difference applies here also. There is almost no difference in the gender summaries in the two assessment processes, apart from a small reversal of the very small gender differences in the upper Year levels. The test model shows a slight advantage for males in upper years, the teacher data a slight advantage for females. The most remarkable feature is the approximate consistency in the gender view, especially when contrasted with the

English/Literacy equivalent. Teachers and test summaries show the same general pattern even though gender is not relevant in the equating process.

An implication of the apparently different trajectories is that in the lower Year level teachers generally report a lower assessment value than does the test for the same student. This difference is reversed in the upper Year levels. It cannot be established from the available data whether this situation is real or an artefact. The test model is influenced by floor and ceiling effects of the individual test and these may account for some of the differences.

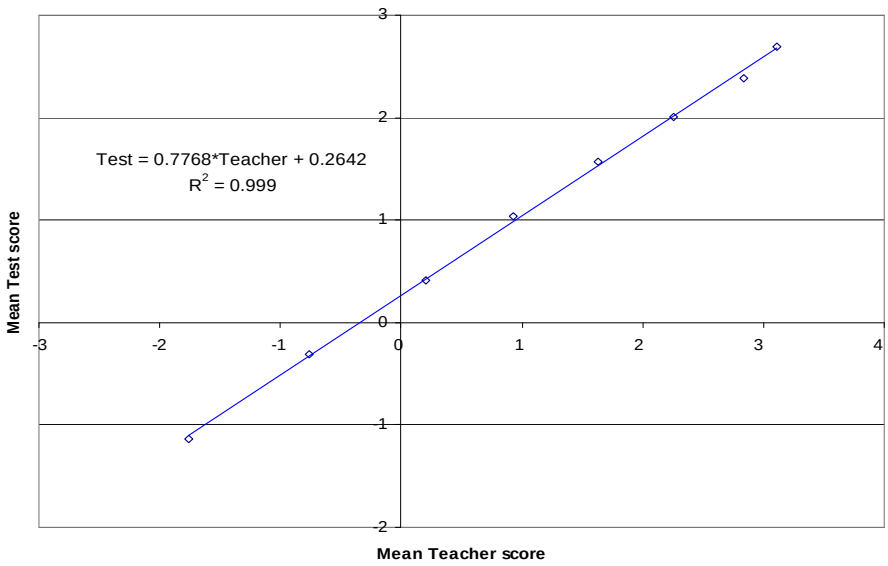
Both the English/Literacy and Mathematics/Numeracy comparisons can be presented by age within Year level. They reflect, generally, the same relationship from the test- or teacher-assessed perspective. However the difference in the trajectories makes the visual comparison by Year level complex. A comparison is made later once the trajectories are equated.

Equating the trajectories

For comparison purpose, the complicating effect of the different trajectories of test and teacher assessments by Year level on the general patterns can be neutralised by equating the trajectories. This is not an equating in the sense applied earlier in the chapter but one of convenience, on the assumption that the trajectories of learning growth with age/Year level should in principle be the same, independent of the particular assessment process. As discussed above it is quite feasible that teachers could consistently under-estimate the learning status of lower Year level students and over-estimate upper Year level students as reflected in the Figures 8.12 and 8.13, particularly in the absence of training and feedback. While in reality it is possible for the trajectories to be quite different, removing this aspect from the data allows a comparison of the degree to which the underlying patterns in the test and teacher assessments reflect the same general phenomena.

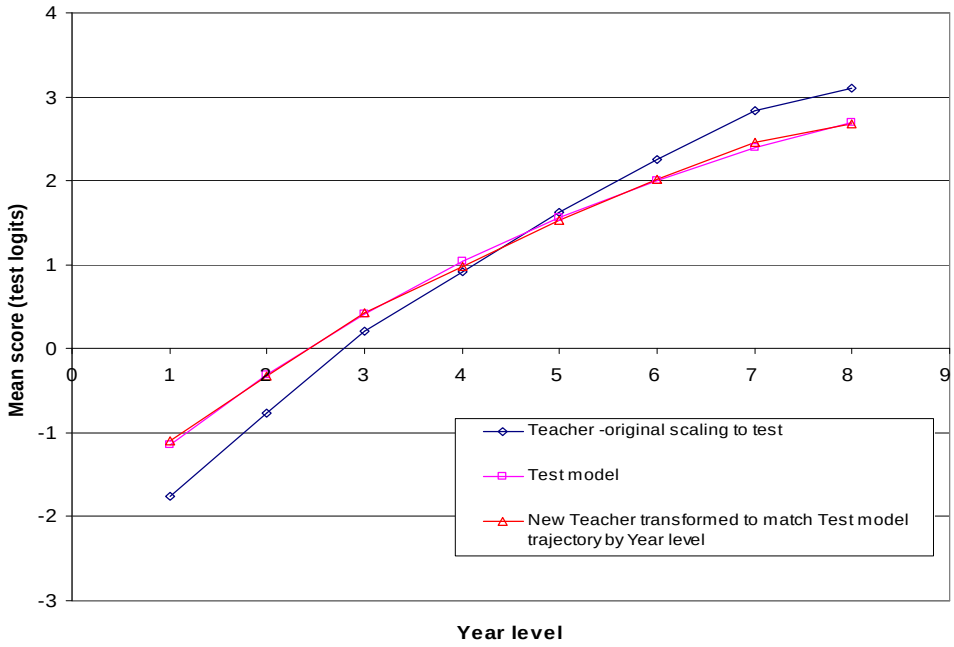
It has been established in the examination of groups of teachers at individual sites above, that at some sites teachers' assessments match the test assessments consistently, that is they are invariant within error. It also established that it is possible for teachers to be consistent assessors relative to the test scale but displaced above or below the expected norm derived relationship and to have a consistent gradient of this relationship with the test scale. The following equating process removes the effect of the difference in trajectory, even if that difference is a real effect. The trajectory equating is achieved by plotting the Year level means for the score values for the teacher and test-model scores. A line (Figure 8.14) is then fitted to the points and this used to transform teacher data (already in approximate test logit units) so that the means at each Year level are the same for both assessment processes. The choice of the test means as the base is for consistency. It does not imply that the test trajectory is the correct or real trajectory.

Figure 8.14 1997 English/Literacy Test and Teacher mean scores at each Year level-Expression to equate means



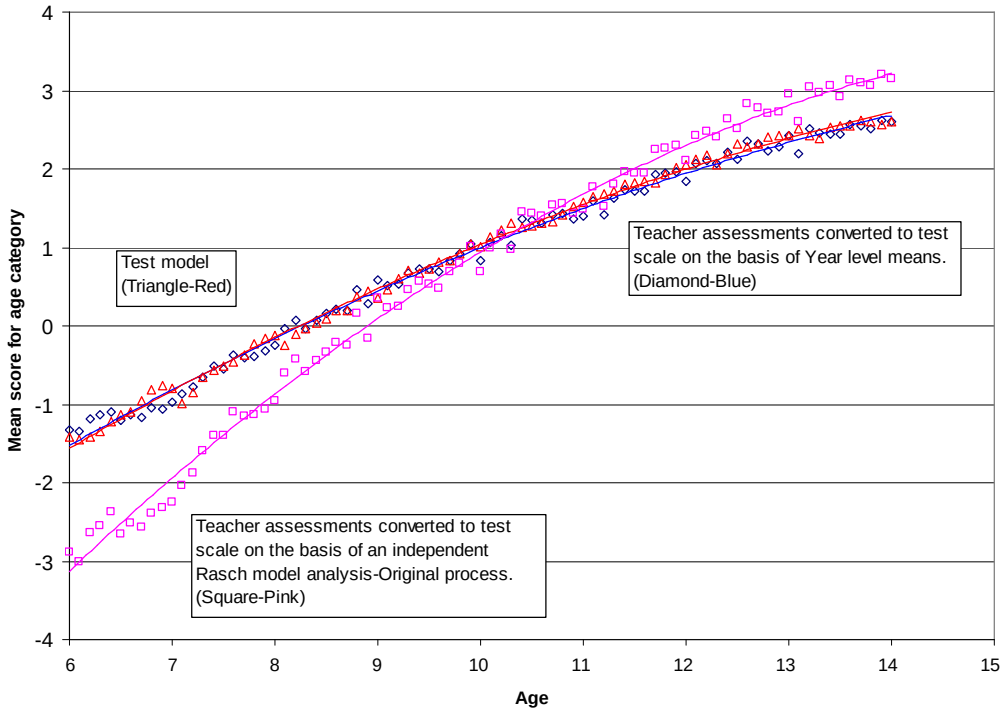
The means for each Year level are equated through linear equating. Through this process the SDs are also equated. The re-scaled teacher data are summarised by Year level to confirm the effect of the additional transformation. The original teacher assessment trajectory in Figure 8.15 is compared with the modified trajectory and the original test model trajectory. The effect of the additional re-scaling has been to make the trajectories identical as intended.

Figure 8.15 1997 English/Literacy-Comparison of original teacher trajectory with the Year level mean re-scaled teacher trajectory and with the Test model trajectory



The result of simple mean equating by Year level of the trajectories for 1998 data, applied for simplicity in lieu of linear equating, is shown in Figure 8.16. In this case SDs are not equated. The data in this presentation are summarised by the age categories in decimal age units, reflecting that the Year level derived transformation adequately transforms the data into an age view. A quadratic line of best fit for the approximately 80 age points is applied in all three cases. The lowest curve is the original teacher data already re-scaled by the Rasch model to the test scale. The age points are transformed (to the diamond points) at the teacher assessment converted to the test scale trajectory. The interwoven curves are the test model and the mean equated teacher data.

Figure 8.16 Effect of Alternative equating processes on Teacher Test assessment comparisons-using Mathematics/Numeracy 1998



Reflection on the impact of the Year level mean transformation to the 'signals' in the data

Before summarising the data in more detail a reflection on the process is useful. Are there some steps in the process of summarising the data that have polluted the teacher or test data so that the results of strong general similarity are guaranteed? Has the development of the test model ensured that the data will match when summarised by age, year level and gender? The data sets are developed independently, at least until the equating steps. The test model is based on estimates of likely trajectories for test data as described in Chapter 6. Curve fitting to vertically scaled tests by Year level, from a number of sources, as illustrated in Chapter 5

produces a consistent generic result for many data sets. The trajectory with age and Year level is a curve with a diminishing growth rate with time. The estimate of IRT test measured student learning growth patterns, are made independently of the teacher data.

The teacher data are fitted to the Rasch model independently of the test data. The original observations by individual teachers generate the data points for each student. Nothing has been done to disrupt or change the natural teacher-observed relationships between strands, ages, Year levels or gender except through systematic transformations. The transformations are based on a Rasch analysis of 1275 (1997) and 2100 (1998) common cases at Years 3 and 5 and then the mean and SDs on the teacher scale transformed to equal those on the test for the common cases. These transformations are then applied to the full teacher data set. As far as the author can see, none of the transformations have contaminated the general trends in the data or ensured that particular relationships should be found. The analysis raises the possibility that teachers may, on average, see student development consistently with test assessments but under or over estimate a student's status depending upon the Year level or age or stage of development.

The transformation of the teacher assessment value to a logit value produces a very similar result when compared with an equi-percentile equating, suggesting that both equating process are approximately equivalent. The equating of the teacher and test scales is limited since it depends on two Year levels only out of 8 (Year 3 and 5), though these are balanced in the central zone of the Year levels of the teacher data. The limited number of common points may influence the relative Year level trajectories of the two assessment processes but will not influence other elements of the data. The effect of the difference in trajectory, even though this may be real, is removed using mean equating for each Year level. When removing the differences due to the apparently different trajectories, the transformation should not affect other general properties of the data. The transformation to equate the means at each Year level is an additional linear transformation of the teacher data. The general correlation coefficients of the teacher and test data sets with each other are unaffected.

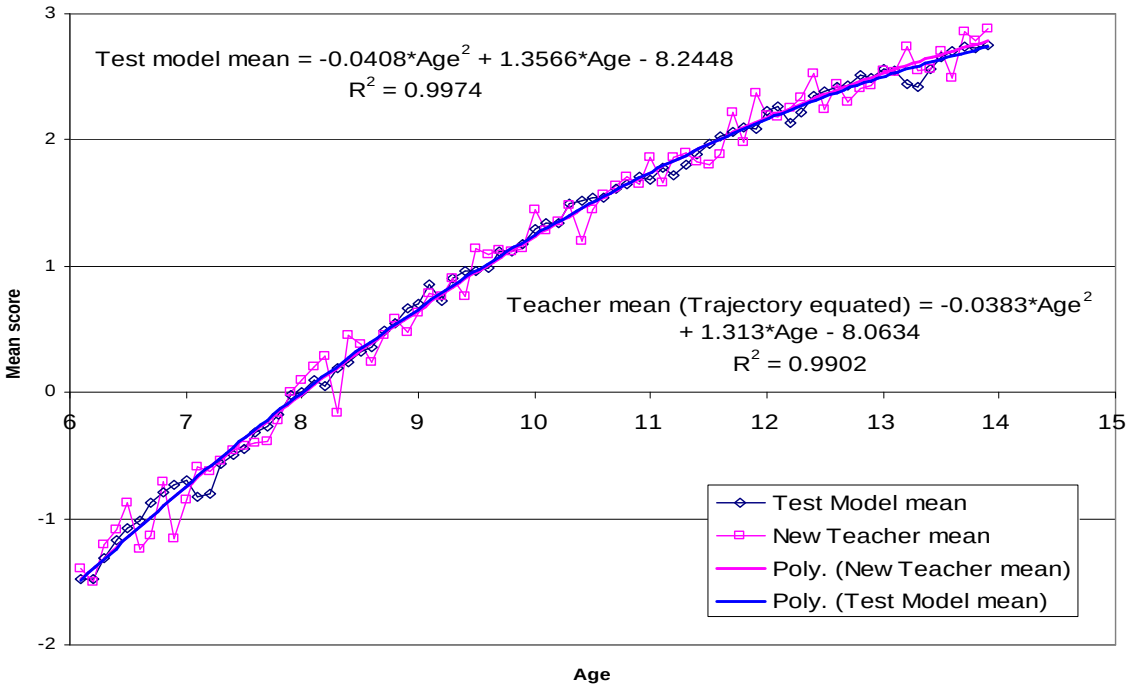
The mean equating ensures that the trajectories match. Thus this aspect, the trajectory of the relationship between test and teacher data, is artificial in the examples that follow. The subsequent comparisons merely provide confirmation that with appropriate transformations the data sets, at the level of mean summaries, can be made virtually identical. However the comparisons that are not directly trajectory dependent are the keys to understanding the degree of consistency in the two assessment approaches.

Key insights from the trajectory transformed data.

1997 English/Literacy

Figure 8.17 reinforces the view that the relationship of age and Year level is such that an equating by Year level (Figure 8.15) will also apply for age. As expected the trajectories of the means by age are virtually identical (as required by the process). In principle however two data sets, when made to follow equal trajectories based on Year level means could show more erratic relationships to the general trend. There are some points for the teacher means (pink squares as points) at each age that vary more widely from the general trajectory than do the points in the test model (blue diamonds). The mean difference from the curve for the teacher means is 0.1 logits over the age range considered. For the test model the mean difference from its curve is 0.05 logits, confirming a closer fit for the test model. The test model means are based on 64,000 cases, the teacher means on 7,900 cases. However the high R^2 values for the quadratic curves suggest that both age relationships with assessment scores are very good fits to the data. The conclusion is that both data sets are very similar. Given that the transformation was based on Year level means (not age) good fit to a similar trajectory implies inherent properties in the teacher data set that follow the same patterns with age as the test model.

Figure 8.17 1997 English-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age



The data in Figure 8.17 can be separated into gender subsets. The result is shown in Figure 8.18. The gender subsets follow essentially identical trajectories, illustrated by fitting quadratic curves. This effect is not determined by the equating of general Year level

trajectories. The teacher data could fit the general test trajectory without the gender subsets of the teacher data matching the test subsets. That the gender trajectories are very similar offers confirming evidence that teacher judgement assessments are remarkably consistent and at a population level (as distinct from individual cases) the general underlying trends and identification of learning status by gender are common to both processes.

Figure 8.18 1997 English-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by gender by age

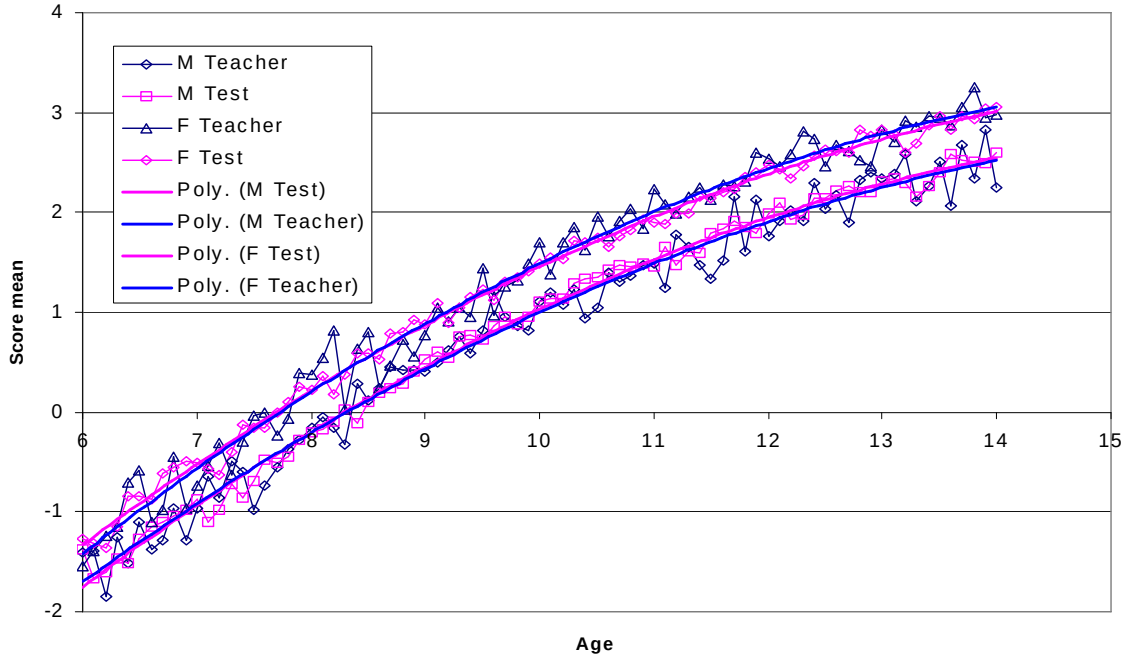
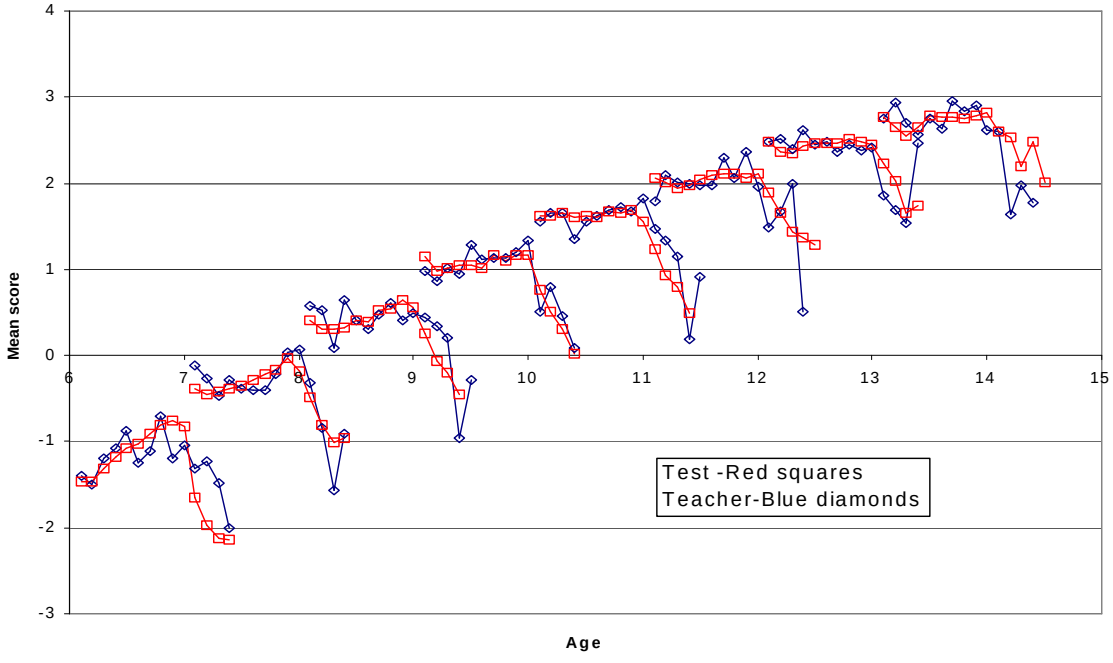
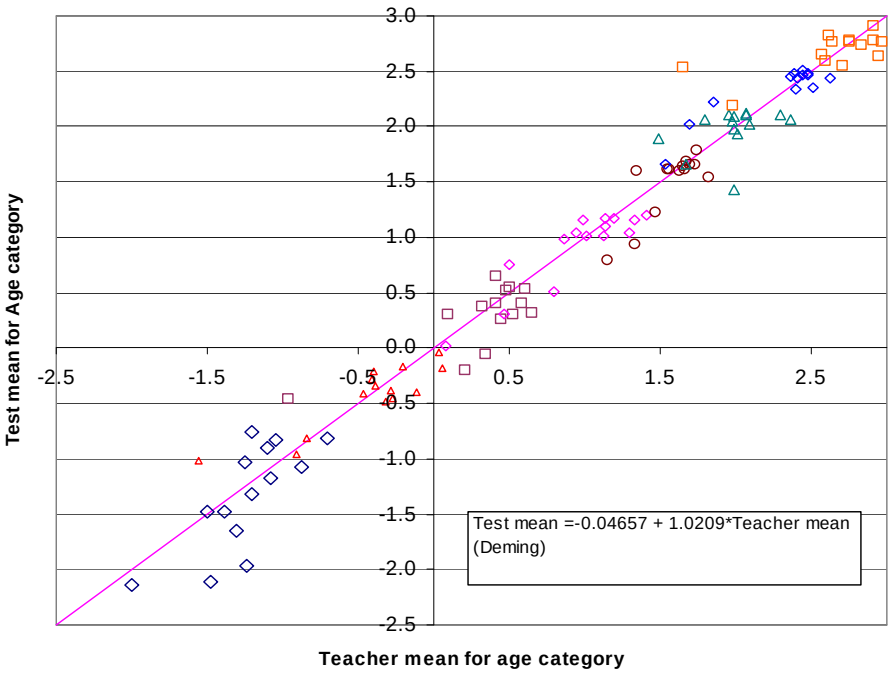


Figure 8.19 1997 English-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age within Year level



When the cases are summarised by age within Year level the relationship of the test means by age to the teacher means is indicated in Figure 8.19. That some points at each year level are coincident is to be expected as a result of the equating of trajectories. However the consistency of the proximity of many points is not a necessary consequence of the trajectory equating. In particular the trailing off of learning status estimates with age above the normal age range for the Year level appear to coincide very well, and are consistent with international data cited in Chapter 5. The age within Year level data are represented in Figure 8.20 plotted along the identity line. A Deming regression of the points results in a regression of Test mean=-0.047 + 1.02*Teacher mean (Standard errors: Intercept 0.035, Slope 0.021) indicating that the assessments are trivially displaced from the identity line and thus confirming a high consistency of assessments by age within Year level under both assessment processes. The mean assessments by age, at the very refined scale of 0.1 of a year of age, are very similar across the range of Year levels 1 to 8. Neither the test model development nor the equating processes have introduced the refined age related characteristic into the data.

Figure 8.20 Plots of points from Test and Teacher assessments from Figure 8.19 (Points are restricted to those within the appropriate range for each Year level)

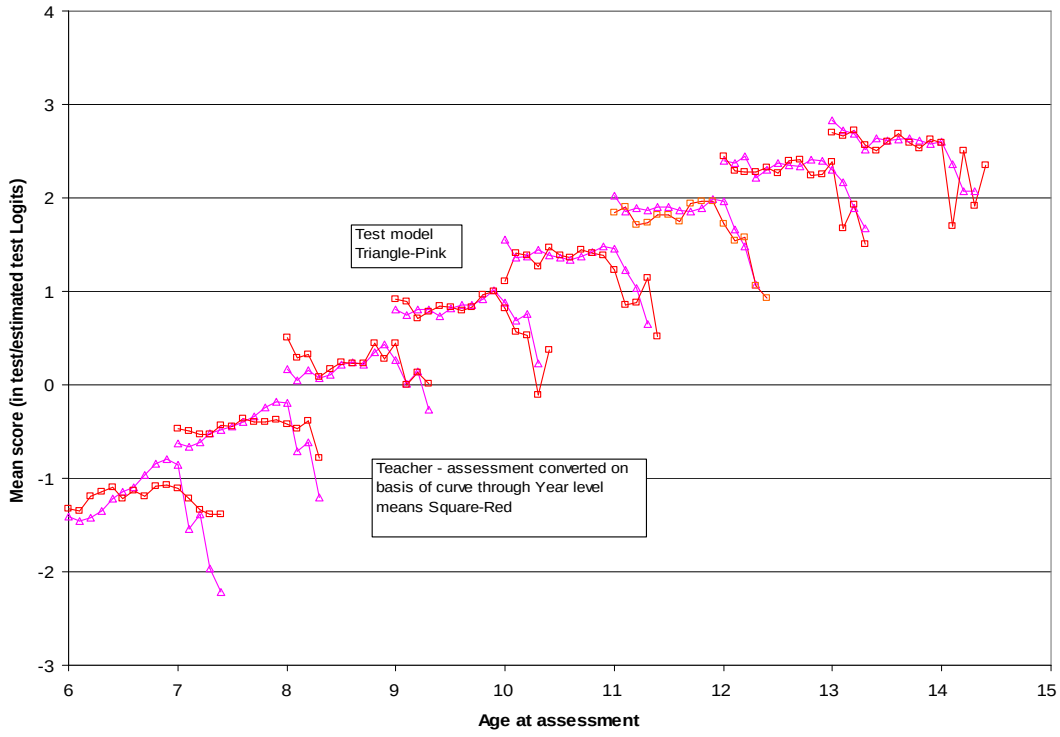


1998 Mathematics/Numeracy

The coalescing of trajectories for the test model and teacher assessments for Mathematics/Numeracy is confirmed in Figure 8.16. The equating of trajectories is on the basis of Year level means only. The same general structure of very closely coinciding data

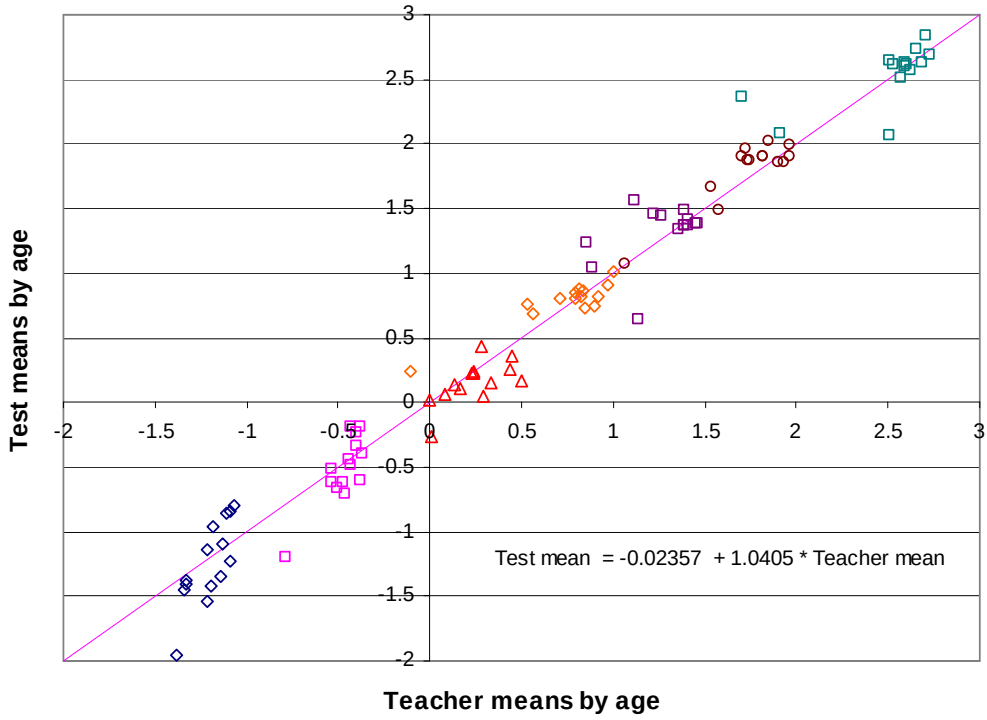
points applies for 1998 data as applied for 1997. The coincidence is less at Years 1 and 2 possibly a result of not adjusting the spread at each Year level. The flatness of the Year 1 was highlighted earlier in Chapter 7.

Figure 8.21 1998 Mathematics-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age within Year level



The relationship of the age data points for each assessment process is illustrated in Figure 8.22. The points cluster along the identity line. A Deming regression of the points has a slope of 1.04 and an intercept of –0.0236 confirming a close fit of the points at each Year level.

Figure 8.22 1998 Mathematics Plots of points from Test and Teacher assessments from Figure 8.21 (Points are restricted to those within the appropriate range for each Year level)



The plot of the gender views from a teacher and a test perspective plotted by age coincide so well that plotting the points of all four components on the one graph generates very close lines. The gender views of mathematics assessments are presented separately in Figures 8.23 and 8.24. Consistent with the English by gender view, the female and male subsets are virtually identical for both assessment processes. However for mathematics, in contrast to the English language examples, both assessment sources indicate almost no difference by gender. This illustrates that the Numeracy test assessments and the mathematics teacher assessments show little difference by gender, whereas the English/Literacy data show a strong gender differences in favour of females consistently in both assessment modes.

There is an indication that teachers in the upper Year levels see the learning status of female students slightly more favourably than do the test data (Figure 8.22), with the corollary that teacher assessments are slightly less favourable for males at the upper Year levels. Overall the two assessment processes produce very similar aggregate results for mathematics indicating that mean teacher assessments and test assessments are almost identical (when trajectories by Year level are equated).

Figure 8.23 1998 Mathematics-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age within Year level: Female students

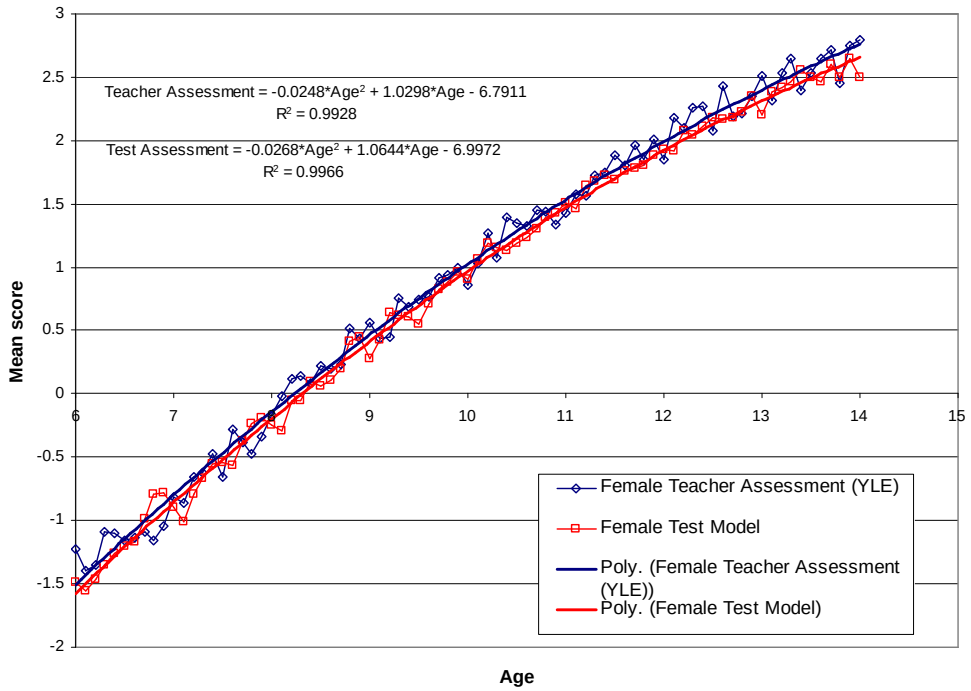
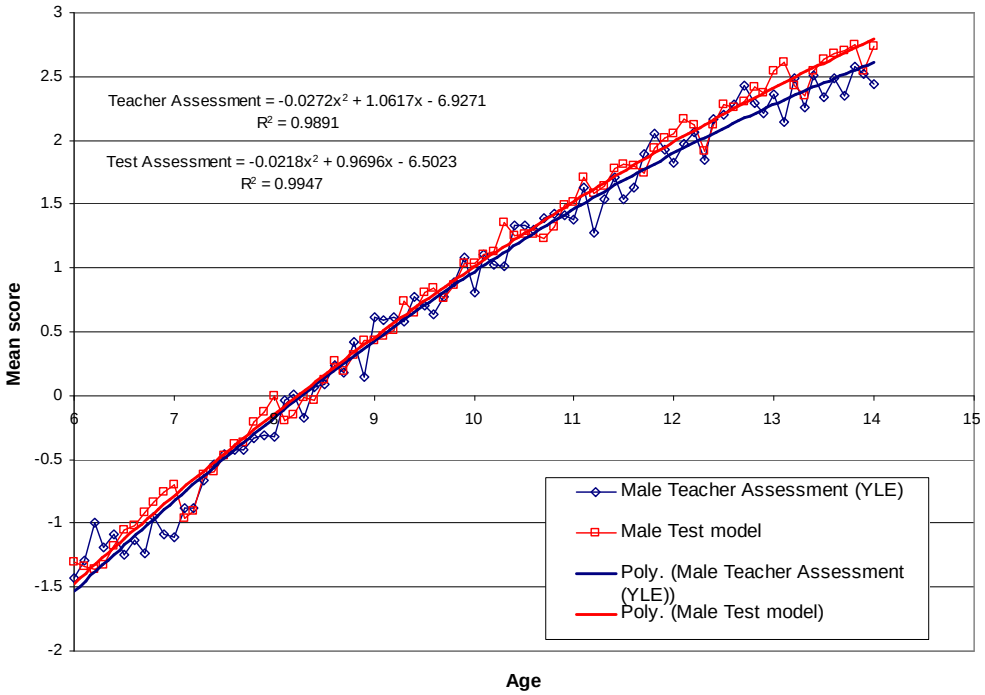


Figure 8.24 1998 Mathematics - Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age within Year level: Male students



Is the variability in assessment alignment a within-teacher or between-teacher effect?

The aggregated data show very strong similarities between test and teacher judgment assessments, particularly when trajectory differences are removed. However there is variation, illustrated early in the chapter and in the school site case studies, in the alignment of individual teachers to the test scale. From the case studies a high correlation of assessments from the two sources can apply even when invariance match rates are low. Even though only just above 50% of assessments are considered as matching, this implies a range of match rates for individual teachers, based on the assumption of a normal distribution of teachers around the mean rate. The data are very restricted (a maximum of 5 cases per teacher) and thus limited in the degree to which conclusions can be drawn. How feasible might it be to improve the alignment of teacher judgment assessments to the test scale?

Some broad speculations based on the 1997 and 1998 cross tabulations of match rates and teacher-test correlations for individual sites can be made. These are explored in Appendix 13. Only sites with more than 5 students assessed are included. This censoring reduces the bias towards high correlations when n is very small. Unfortunately it also eliminates small schools from the analysis biasing the analysis towards larger school sites. The match rates and correlations for the included sites are calculated and ascribed to the estimated number of teachers at the site and tabulated in Appendix 13. While assuming all teachers at a site are equal removes the between teacher differences at a site, it allows a rough estimate to be made of the proportions of teachers who have varying mixes of matching to the test and varying degrees of correlation to the test scale across the set of schools included. Over the full set of students with teacher and test assessments (1275 in 1997 and 2105 in 1998) an estimated 700 teachers are potentially included. Limiting the analysis to sites with more than 5 students reduces the number of teachers included to about 600. Summarised statistics from Appendix 13 are tabulated as shown in Table 8.8. Cut points for low/high are above 0.7 for correlation coefficients and above 0.4 for degree of match of assessments.

Table 8.8 Estimates of the percentage of teachers in categories of correlation with the tests cross-tabulated with the rate of match to the test-1997 and 1998 data combined

	Low Match	High Match	Total
Low Correlation	10%	30%	40%
High Correlation	10%	50%	60%
Total	20%	80%	100%

High correlation: 0.7 or above, Low correlation 0.6 or less. Coefficients rounded to one decimal.
High match: 0.5 or above, Low match 0.4 or below. Match rates rounded to one decimal.
N estimated to be 600 teachers in Years 3 and 5 for 1997 and 1998 combined.

Most site match rates are in the range 0.5 to 0.7 (see Figures 8.8 and 8.9). Only 20% of teachers are estimated to be 0.4 or below, in 0.1 match rate categories. Then, as a general estimate, about 20% of teachers have low match rates. About half of these are estimated to also have low correlation rates. The case studies presented earlier show it is possible to have moderate to high correlation coefficients and still have low match rates due to scale displacement. Accordingly, about 10% of teachers are estimated to have assessments with both low match rates and low correlation coefficients. The assessment approaches of this set of teachers would need to be better understood. They would be amongst the highest priority in developing strategies to improve teacher assessment calibration to the test scale. It might be established that a proportion of these teachers is unable to discriminate learning changes at all and thus cannot be calibrated to the test scale. On balance the speculative data offers optimistic possibilities for the improvement of teacher calibration to the test scales. Strategies that might achieve this include coaching, individual feedback on their current relationship to the test scale and specific training about the meaning and value of the scale in recording progress and intervention options. That consistency within a site can be established (even if miscalibrated) offers evidence for the potential to improve individual teacher calibration.

Summary

This chapter compared two independent methods of student assessment in two learning areas, to investigate the degree to which they appear to arrive at similar assessment results for individual students. The analysis of individual cases at Years 3 and 5 was expanded to compare samples of students assessed by teachers using a standard assessment framework compared to a model of test results for Years 1 to 8. In these comparisons the test and teacher-assessed 'samples' of students are notionally from the same Year level populations, independently sampled rather than being specific students with assessments from both sources.

Comparing teacher and test assessments for the two learning areas presumes both methods are assessing essentially the same skills and behaviours. The common patterns by gender and age within Year level within learning areas established in the analysis, offer support for the validity of the comparison. The clear differences between the results for English and mathematics across learning areas and the similarity of the results between methods within learning areas provide evidence that teachers, on average, produce aggregated summary level assessments that are measurably equivalent to the results from tests.

The links between the two assessments scales for each learning area that allow teacher and test assessments of student learning status to be placed on the same scale were established through a general transformation of the teacher assessments, using a Rasch model analysis.

This equating of scales is based on the assumption that measures of central tendency of the assessments of a large population of teachers are the best basis for equating the scales. Alternative methods of equating based on a series of specific behaviours and skills, a criterion rather than normative basis for equating, might establish a different equating relationship. Such a process was not feasible with this historic data. If such a criterion process were applied, the author assumes the result would affect only the placement of the control lines for the match zone, influencing which cases were deemed to match, not the general degree of match.

Once the assessments scales were equated, the two methods of assessment were compared for degree of match. On the basis of the estimates of measurement error for the two assessment processes, just over 50% of the assessments of students in each learning area are deemed as measurably invariant (i.e., within error) in Years 3 and 5.

The errors of measurement in both assessment processes reinforce that all assessments ere likely vary from the inferred 'actual' status by quite an amount. This illustrates the risk of using one-off assessments – whether by teacher or test. Regular re-assessment is one of the tasks teachers routinely do as part of their normal practice, except that a common vertical scale is rarely applied. Regular re-assessment by test processes on a vertical scale is also feasible, in the form of computer adaptive testing, but at a considerable additional cost.

There is evidence from the site case studies that a common assessment culture, to consistent criteria across a number of teachers, can apply within a school. This consistency applies even though the assessments themselves are not regarded as matching on the basis of the norm-developed translation of teacher assessments to the test scale. The general options of mis-match were speculated upon in Chapter 4. The case studies support that speculation. The quantitative process for articulating the differences between a teacher's set of assessments and the test scale provide a potential basis for improving the calibration of the teacher to the test scale.

Even though the student assessments may not meet the stringent criteria for matching, the correlation of the order of the students on the two assessments axes is often high. When approximate consistency of order applies the relationship of teacher to test assessments can be understood through the gradient and intercept of the comparison plots. Assuming a symmetric regression that allows error on both axes, the slope when near one indicates a consistent spacing of the assessments on both axes. Gradients markedly greater or less than one imply compression (or expansion) of the scale on one of the axes relative to the other. Zero or vertical gradients indicate no relationship of the two scales. Negative gradients imply a reverse order of students in the two processes.

The other aspect of difference, the intercept, indicates (based on its sign) that one assessment scheme is assessing above or below the other. A gradient close to 1 but an intercept markedly different from 0 implies a systematic displacement of the scales. This was illustrated in some of the cases studies and was consistent across multiple teachers at a site. Teachers in this situation are assessing the students with consistent order and spacing on the teacher scale as applies on the test scale but with systematically higher or lower values. Since the values on the scale have specific meaning in terms of what students can and cannot do, the displacement negates the interpretive value of the scale. The consistent order but displaced relationship to the test scale indicates a need to clarify the link between the scales. The consistency implies it would be feasible to recalibrate teachers to an interpretation of the level scale that is more closely aligned to the test scale. An implication of the consistent displacement is that fewer student assessments are seen as matched even though the teachers and the test are assessing consistently relative to each other.

The measurement characteristics of a test are fixed by its design. It works approximately the same in all applications. The teachers on the other hand are not automatically aligned to the test scale. Each alignment is a personal calibration. It would appear from the case studies that teachers could be aligned with each other and consistently to the test scale at varying degrees of displacement. A closer match would appear feasible through coaching, training and feedback from regular personal test comparisons for each teacher. Based on the estimates of individual teachers' correlations with the test scale, very few would not be able to be linked to the test scale. A universal alignment would seem possible for many teachers, subject to a prior step of re-examining and refining the scale structures.

The teacher assessments appear to show different perceptions of the trajectories of learning growth with age/Year level relative to the test even when the teacher scale unit is transformed to the test scale. This is an implication of the apparently differing trajectories of the test model compared with the transformed teacher data. The differences in the trajectories, if real, imply that teachers, relative to tests, underestimate skill levels for younger students and over-estimate skills for older students. More likely, however, is that the difference is a combination of inadequate modelling and a distortion of teachers' assessments due to the calibration issues discussed in the chapter. Resolving whether there is a systematic variation in the teachers' perceptions of the scales requires more teacher and test data at each Year level and a better study design. Such a study would be feasible in Victoria where data from both assessment processes are available at Years 3, 5, 7 and 9 at least.

Equating the trajectories for teacher and test assessments (i.e., eliminating the systematic differences) shows that the general features of the mean learning status as reported by teachers and estimated from tests are very similar. The gender analyses are consistent

between test and teacher assessments within learning areas. The mean learning status by age structures within a Year level are very similar with both showing a characteristic tail for over-age students. It is most unlikely that teachers consciously consider age and gender aspects when making their learning status judgement. There is a small over-estimation by teachers in favour of girls in higher Year level mathematics relative to test assessments.

In summary, an adequate view of population learning patterns by Year level and age and learning area can be obtained from systematic teacher assessments approximately standardised through a level structure. Whether a time series of an individual student's assessments can be used to monitor learning growth is less clear. Were it possible, the spin-off benefit of this to personalised management of student learning development would be high, through consistent and meaningful interpretations of what a given scale value implies about skills to be developed next. How teacher assessment for individual students might be integrated into an improved learning support and management system is addressed in the concluding chapter.