Chapter 6: South Australian test data for 1997 and 1998

> A measure of growth is based on measures of status on three or more occasions, obtained either by averaging two or more 'gains' or by modelling growth (curve fitting).
>
> Masters, Rowley, Ainley & Khoo, 2008, p. 16.

The previous chapter explored trajectories of learning. One purpose was to develop a model of test assessment means over a wide range of Year levels to provide a basis for comparison with teacher judgment assessments of students. Teacher judgement assessments, described in the next chapter, provide a consecutive Year level view from Year 1 to Year 8. Test assessment data on the other hand only exist for Years 3 and 5 in South Australia in 1997 and 1998.

To provide a comparable test assessment view over Years 1 to 8, a model of consecutive Year level test assessments is developed in this chapter, extrapolated from known data for Years 3, 5, 7 and 9. These data are drawn from tests for the same student populations, as close to the period of teacher judgement assessment as possible. Data for non-tested Year levels are imputed at a student level using the characteristics of learning growth with Year level and age illustrated in Chapter 5. The extent to which the resulting models represent the real situation is contestable, as for all models. The purpose of the models in this broad analysis is to provide a basis for approximate comparisons of the results of the alternative assessment processes. Accepting the limitations of the models, do teacher and test assessment approaches represent learning development in sufficiently equivalent ways?

The Basic Skills testing program (BSTP) commenced in South Australia with trials in 1994 and full cohort testing for Years 3 and 5 implemented in 1995. A brief history of the implementation of the BSTP is covered in Chapter 3 and is available in more detail in Hungi (2003). These tests have already been part of an extensive publicly reported analysis (Hungi, 2003) covering the years 1995 to 2000. The mean Year level scores for Years 3, 5 and 7 from the tests along with the individual student scores enable a speculative model of individual student scores from Year 1 to Year 8 to be developed. This model, while an imperfect substitute for actual data, provides a basis for a comparison with teacher judgement assessments for the same Year levels. The two data sets are compared later in Chapter 8.

Literacy and Numeracy Tests

The Basic Skills test had two main parts; a Literacy section and a Numeracy section, with subscales within each section. Test items were vertically scaled through common items in the

two Year levels. Student responses were analysed using the Rasch model. Scores were transformed from the original logit scores and reported in a range from 0 to 99, to one decimal point. This thesis uses the original logit scores. As reported in Chapter 3, the South Australian tests were the same tests as used in New South Wales schools for the same years. The NSW Department of Education developed the tests and provided South Australia's results.

The Department of Education and Children's Services (DECS), known as the South Australian Department of Education and Training in 1997 and 1998, approved access to data in February 2005 (Appendix 1). In all DECS made available: individual test records for Years 3 and 5 for the 1997 and 1998, Year 7 records for 2001 (the first year for Year 7 testing), 2002 and 2004. It also made available teacher judgement assessments for 1997 and 1998, reported in Chapter 7. Table 6.1 indicates the number of students included in the main test data files used.

Table 6.1 Students in the Basic Skills Test Program (BSTP) included in data analysis

	1997	1998	2001	2002
Year 3	12437	12794		
Year 5	11973	12471		
Year 7			12873	12930

Test measures of the students in Years 3 and 5 in Literacy in 1997 and Mathematics in 1998 are presented for general comparisons with teacher judgement assessments for these same cohorts. Detail of individual items and the performance of specific items is not a focus. Students' scores are the main interest.

Rasch model analysis of the Literacy and Numeracy tests

To reconcile files provided, to be assured that their structures were fully understood and to re-estimate the error of measurement for each student, a Rasch model analysis using Winsteps was applied to each of the data sets. This was a repeat of the original analysis by the NSW Department. The original data files provided to the author included the Literacy and Numeracy scores for each student on a common scale for Years 3 and 5 in logit form as well as the item responses for each student. They did not include SEs or fit statistics. Details identifying common items in the Year 3 and Year 5 tests were also missing. As a result the Rasch analysis was run for each Year level independently[24] and then crosschecked for consistent results with the originally supplied student logit scores.

[24] Logit scores on the common scale were already known and the only missing details were error of measurement estimates and fit statistics. A request for further information from the SA was deemed

Fit and measurement statistics obtained are tabulated in Tables 6.2 and 6.3. These are presented to confirm the general adequacy of the test. Results from this reanalysis were compared with the original summary scores and the recreated errors of measurement, infit and outfit values added to the record for each student. This was done to establish the individual errors of measurement to be used later in comparisons with teacher judgement assessments.

Table 6.2 Summary of Winsteps Fit and Measurement Statistics, Literacy 1997

Items	N	Measure (mean)	Model Error	SD of Measure	SD of Error	Reli-ability	Separ-ation	Real RMSE	Adjust-ed SD	Infit MS	SD of Infit	Outfit MS	SD of Outfit
Test Y3	58	0.00	0.02	1.00	0.00	1.00	40.80	0.02	1.00	0.99	0.12	0.98	0.24
Test Y5	83	0.00	0.03	1.25	0.01	1.00	45.11	0.03	1.25	0.99	0.11	0.96	0.23
Students													
Test Y3	12437	1.03	0.37	1.30	0.11	0.91	3.24	0.38	1.24	1.00	0.12	0.98	0.33
Test Y5	11972	1.42	0.33	1.22	0.08	0.92	3.45	0.34	1.17	0.99	0.15	0.96	0.37

Students	Above 1.3 Infit MS	Below 0.7 Infit MS
Test Y3	1.8%	0.1%
Test Y5	2.9%	0.7%

Table 6.3 Summary of Winsteps Fit and Measurement Statistics, Numeracy 1998

Items	N	Measure (mean)	Model Error	SD of Measure	SD of Error	Reli-ability	Separ-ation	Real RMSE	Adjust-ed SD	Infit MS	SD of Infit	Outfit MS	SD of Outfit
Test Y3	32	0.00	0.02	1.19	0.00	1.00	48.52	0.02	1.19	0.99	0.10	1.00	0.18
Test Y5	48	0.00	0.02	1.25	0.01	1.00	48.57	0.03	1.25	0.99	0.07	0.99	0.15
Students													
Test Y3	12794	0.91	0.49	1.25	0.11	0.84	2.27	0.50	1.14	1.00	0.18	1.00	0.50
Test Y5	12471	1.03	0.39	1.10	0.09	0.87	2.55	0.40	1.03	1.00	0.15	0.99	0.44

Students	Above 1.3 Infit MS	Below 0.7 Infit MS
Test Y3	5.6%	2.0%
Test Y5	3.3%	0.8%

unnecessary since the individual student errors of measurement from the individual Year level analyses were assumed to be similar to those found in the common analysis.

Comment on Tables 6.2 and 6.3

The mean of the Year 3 Literacy student scores was 1.03 logits. Effectively the mean difficulty of the test items (0.0) was 1.03 logits easier than the average learning status of the students taking the test, meaning that the test was well targeted, certainly not too hard for the majority of the students. While the original linked analysis placed Year 3 and Year 5 items on a common scale, the relative item placements and spacings of the Year 3 items would be expected to vary only slightly between the linked analysis and the unlinked analysis.

The Year 5 Literacy test was also easier than the average learning status of the students taking the test by 1.42 logits, making it relatively easier for the Year 5 students than the Year 3 test was for the Year 3 students. The test was not too hard for the majority of students. The recreated Year 3 and Year 5 scores from the unlinked analysis scores for each student differed systematically from the original logit scores provided from the original linked analysis. Author-derived Year 3 1997 literacy scores for each student were consistently about 0.67 logits above the original combined Year 3-Year 5 scores provided by the NSW Department, developed using an across-Year level vertical scale.

Year 5 student scores were consistently 0.13 logits below the linked scores. The consistency of relationship at an individual student level between the combined and separate analyses indicates both analyses obtained equivalent scores for students, and that the re-estimated errors of measurement for each student could be assumed to be equivalent to those obtained in the original NSW linked analysis, allowing them to be used in a confidence interval comparison at a later stage in the analysis.

Standard deviations for the distributions of scores were slightly lower for the unlinked analysis than for the linked analysis (1.30 compared with 1.36 for Year 3, 1.22 compared with 1.24 for Year 5). This difference is due to the wider range of student scores in the combined analysis. That the difference is so small indicates that the range of scores for each Year level separately is almost identical to the combined range. Similar patterns applied for Numeracy in 1998 (Table 6.3), with the standard deviations increasing slightly in the linked analysis.

For subsequent summaries and analyses the original student scores from the linked scales analysis were used, with re-estimated errors of measurement and fit statistics added to each student record.

The mean score differences (that is growth) for the original logit scores were 1.19 logits (Year 3 mean 0.36, Year 5 mean 1.55) for Literacy in 1997, and 1.21 logits (Year 3 mean 0.13, Year 5 mean 1.34) for Numeracy in 1998. Hungi (2003) applied an analysis with a more sophisticated equating and linking process than generally applied by NSW analysts. He established that a linked analysis over the calendar years (1995 to 2000) varied the original

mean scores at each Year level and for each calendar year by up to 0.05 of a logit and the resultant growth estimates from Year 3 to Year 5 by up to 0.1 of a logit (see Table 6.4 and 6.8). This variation is equivalent to about $1/5^{th}$ of a year's learning and adds an additional tolerance consideration when test and teachers judgement assessments are compared.

The fit of Literacy items in 1997 and Numeracy items in 1998 to the Rasch model was good as shown in Tables 6.2 and 6.3. Infit mean square values for the items were close to 1.0 with none outside the range 0.7 to 1.3. The tables indicate the percentages of students with infit values above and below the values of 0.7 and 1.3. Cases below 0.7 are negligible except for Year 3 Numeracy where 2% are estimated to overfit implying a small degree of item dependency for these students. Misfitting students (above 1.3 Infit values) are between 2% and 3% of cases, except again for Numeracy in Year 3, where almost 6% of students misfit.

Item reliabilities are consistently 1.0 (due to the very large N of students tested). Reliability from a student measurement perspective is less for the Numeracy test (0.84, 0.87) than for the Literacy test (0.91, 0.92). The mean model error of measurement for students is of the order of 0.3-0.4 logits (0.49 for Year 3 Numeracy) implying an error in estimating individual student learning status of up to 8 months learning development. This error is greater for Year 3, most likely reflecting the complexities in measuring numeracy skills for students with low numeracy and most often low reading skills with a self completed paper and pencil test. The high percentages of students with infit values above 1.3 infit mean square in Year 3 (Table 6.3) supports this explanation.

At this stage in the data development, of the order of 12,000 records for Year levels 3 and 5 (as shown in Table 6.1) are available with student scores, errors of measurement and fit statistics. These records serve two purposes. About 1000 cases per year level have a potential match with the teacher judgement assessments to be taken up in Chapter 8. The second purpose is to contribute student cases to the development of model data sets using the understanding of trajectories of learning from the previous chapter to impute values for notional students in Year levels 1 to 8. This is done to provide a comparison with the teacher judgement assessments. If students had been assessed by tests at all year levels what might that data have looked like?

The trajectory of Literacy test scores

To aid in the estimation of the trajectories of learning as Year level increases, a wide range of South Australian data are reviewed. These data are considered as part of the process to select data points for the mean scores in Literacy at each of the tested Year levels. A curve is fitted to these points using the Gompertz expression as described in Chapter 5. This trajectory then becomes a framework for estimating the means for missing Year levels. Data for typical

students at each missing Year level are then imputed. This is done by adjusting the scores for a random sample of students drawn from the tested Year levels (3, 5 and 7) so that the mean scores approximately match the framework. Adjacent cohorts (3 and 5, 5 and 7) contribute equally to the samples for Years 4 and 6. Cohorts below Year 3 are 'stretched' to ensure that both the overall means and the means by 0.1 of age follow the framework trajectory. This is achieved by including the age at testing as well as the Year level for each imputed student. The age at testing values for each student then allow a more general set of summaries of the model data to be made.

Table 6.4 summarises the mean scores in Literacy for SA students by tested year level from 1997 to 2004 as well as NAPLAN data for 2008. Data for 2003 were not in a form that could be readily summarised and are omitted. Four perspectives of the data are provided, two cross-sectional and two longitudinal. This is done to establish that all four perspectives generate essentially the same curve of learning development with Year level and age, confirming the Hilton and Patrick (1970) finding that the general change from testing period to testing period was similar whether the group of interest was cross sectional or longitudinal-not matched (the cohort not adjusted for losses and gains).

The first cross-sectional block of Table 6.4 (1997 as an example) presents the original SA cross-sectional values at Year 3 and Year 5 (and Year 7 in some cases). The range of scores within a Year level is wide (0.18-0.61 at Year 3, 1.38-1.79 at Year 5). Growth values are in a narrower range (1.03 to 1.23 logits for Year 3 to 5, 0.70 to 0.89 for Year 5 to 7).

The second cross-sectional block (1995, Hungi adjusted) is the result of Hungi's re-scaling based on a multi-linked analysis. The mean values at each Year level estimated by Hungi are not strictly comparable to those values originally derived for each calendar year. The value of 0.31 (1997 Year 3 as adjusted by Hungi) on the common item scale for example, may not be positioned at exactly 0.31 on the original scale, since Hungi refined the scale to be better calibrated across the testing years 1995 to 2000 than was originally developed in NSW. An implication of the Hungi analysis is that the NSW item scale had the potential to vary from calendar year to calendar year. The growth estimates at the right side of the table represent differences between scale values and should be more comparable even if not in identical units.

The third block presents a longitudinal view (cohort wave identified by the calendar year at Year level 3) showing the values for the Year 3 cohorts at successive 2 yearly intervals. As an example the Year 3 value in 1997 is related to the Year 5 value in 1999, two years later.

The fourth perspective is also longitudinal, similar to the third, but uses the re-calibrated Hungi values to indicate growth values based on the same cohorts two years apart.

Table 6.4 Literacy – Mean scores by Year level and Testing Year

	Year Level	3	5	7	9	Growth	Growth	Growth
	Average age	**8.6**	**10.6**	**12.6**	**14.5**	**3 to 5**	**5 to 7**	**7 to 9**
Cross-sectional	1997	**0.36***	**1.55***			1.19		
	1998	0.18	1.42			1.23		
	1999	0.44	1.47			1.03		
	2000	0.30	1.38			1.08		
	2001	0.45	1.60	2.30[25]		1.15	0.70	
	2002	0.39	1.61	2.41		1.22	0.80	
	2004	0.61	1.79	2.68		1.18	0.89	
	Mean	0.39	1.54	**2.46***		1.15	0.80	
	1995 (Hungi adjusted)	0.38	1.36			0.98		
	1996 (Hungi adjusted)	0.30	1.48			1.18		
	1997 (Hungi adjusted)	0.31	1.57			1.26		
	1998 (Hungi adjusted)	0.22	1.50			1.28		
	1999 (Hungi adjusted)	0.38	1.31			0.93		
	2000 (Hungi adjusted)	0.18	1.29			1.11		
	Mean	0.30	1.42			1.12		
Longitudinal	1997 Cohort wave	0.36	1.47	2.30		1.11	0.83	
	1998 Cohort wave	0.18	1.38	2.41		1.19	1.04	
	1999 Cohort wave	0.44	1.60			1.16		
	2000 Cohort wave	0.30	1.79	2.68		1.49	0.89	
	Mean	0.32	1.56	2.46		1.24	0.91	
	1995 Cohort wave (Hungi)	0.38	1.57			1.19		
	1996 Cohort wave (Hungi)	0.30	1.5			1.20		
	1997 Cohort wave (Hungi)	0.31	1.31			1.00		
	1998 Cohort wave (Hungi)	0.22	1.29			1.07		
	Mean	0.30	1.42			1.12		
NAPLAN	NAPLAN SA 2008 Reading (estimated logits)	0.40	1.51	2.30	**2.89***	1.11	0.79	0.59
	NAPLAN SA 2008 Reading (reported scores)	*(400.5)*	*(477.9)*	*(533.5)*	*(575)*	*(77.4)*	*(55.6)*	*(41.4)*

* Bold values signify values used in estimation of Gompertz model parameters.

The final block of the table indicates the reading data for SA's 2008 NAPLAN tests. This is not the same as the more general Literacy applying in 1997 but is assumed to be a reasonable indicator of the general trajectory. These data are tabulated to provide an additional influence on the model trajectory for Years 7, 8 and 9 developed below. The original scores are provided, along with a conversion on an estimated basis to logits. Growth from Year 3 to Year 5 over a number of years is estimated to be approximately 1.12 logits based on the grand mean of Hungi's estimates. The growth of 77.4 NAPLAN units is used to estimate an approximate conversion factor of 70 units to 1 logit. On this basis, after setting the NAPLAN

[25] The estimates for the means at Year 7 are derived from the original files provided. About 300 cases were omitted for 2001, 50 for 2002. These cases had the lowest possible scores for Literacy but high scores for Numeracy. It is assumed that these were cases with no data for the specific test and allocated minimum scores in the original analysis. The author's summary omits them.

value at Year 3 as 0.4 logits (to fix it close to the Year 3 1997 Literacy value), approximate estimates in logits can be made for the reading means at Years 5, 7 and 9. The resultant growth values are comparable to the other growth trends in the table[26].

In summary, the growth in mean literacy learning over a mix of years and cases when viewed as cross-sectional is very similar for the original SA data and for the Hungi re-scaling. For longitudinal growth, the Hungi re-scaling is also similar to the cross-sectional growth. The growths in the longitudinal view of the original data are also similar, although the mean is possibly skewed by the 2000 Cohort wave value. This is illustrated graphically later in the chapter and is possibly an artefact of less accurate across-calendar year equating that Hungi has addressed in his re-equating of the scales. Overall this extensive analysis of the change of scores from one tested Year level to the next tested Year level establishes that the patterns of group mean growth in scores are essentially the same whether a cross-sectional or longitudinal view is used. The period of interest (1997) very conveniently approximates the mean values for all the cases examined. The 1997 values plus the NAPLAN extension to Year 9 are used in the next section to develop a framework for a model of literacy growth in South Australia in 1997.

Developing a model for the test trajectory of learning – Literacy.

The model is developed in two stages. Initially a general model for the mean learning status at each year level is developed: the framework. In the second stage, data points for individual students at each missing year level are created, based on actual student data from Years 3, 5 and 7. A number of assumptions are made in the model development.

Patterns of mean learning development by Year level are assumed to be similar over different calendar years. While these patterns vary, evidence from Chapter 5 (and Tables 6.4 and 6.8) indicates that these patterns are consistent enough to provide a trajectory framework for group means. It is assumed the location of the mean for an intermediate Year level (Year 4, Year 6) can be placed on a smooth curve describing the trajectory, accepting the Year level units as equally spaced time units on the X-axis. In the absence of actual SA test data describing the trajectory of literacy learning leading up to Year 3, it is assumed that the means for Year 1 and 2 can be placed on a smoothed trajectory, using a Gompertz expression as outlined in

[26] The average age for each cohort has increased in the period from 1997 to 2008. Based on estimates from the annual age and Year level census bulletins (ABS 4221), the average age at July 1 has increased by about 0.15 of a year of age at Year 1 and by 0.1 at Year 7. The testing period has also shifted to earlier in the school year (August in 1997 to May in 2008), meaning that the age at testing has not varied much over this period.

Chapter 5 to estimate the trajectory. As illustrated in that chapter, when the NAPLAN model is compared to a number of US normed tests, a steeper growth curve in the early years of school is found for the US data. For Literacy the Gompertz model trajectory for SA data therefore may not be steep enough. The means for Years 1 and 2, derived from the Gompertz curve fitted to the data in Table 6.4, may be conservative in the degree of learning growth they indicate.

Even if conservative, a steeper trajectory for learning applies in the early years relative to later years and the within-Year level by age curve has a steeper gradient in the lower years. Noting this age effect described in Chapter 5, the spread of scores by age within Years 1 and 2 needs to be increased to reflect this effect. An extensive record matching process to add dates of birth to the student score files is required. The age for each student at testing can then be calculated. While Year 7 scores for 1997 do not exist, it is assumed that the records of approximately 26,000 students (see Table 6.1) in 2001 and 2002 are indicative of the general spread of student scores at this higher Year level. These records are used to add Year 7 and Year 8 cases for the model.

Finally it is assumed that the Year 7 mean scores, combined with the general estimates from the SA NAPLAN data for Year 9, can be used as points in the model to influence the trajectory fitted for higher Year levels. Assuming that the growth from each tested Year level to the next is consistent over testing periods is only partly valid. The larger the test population, the more likely the growth in learning remains consistent. England and US national data reviewed in Chapter 5 show very small variation in mean scores over time. As the unit of analysis becomes smaller, at a school or classroom level, much greater variations in the scores are observed. These variations appear to balance each other out in the aggregated summaries.

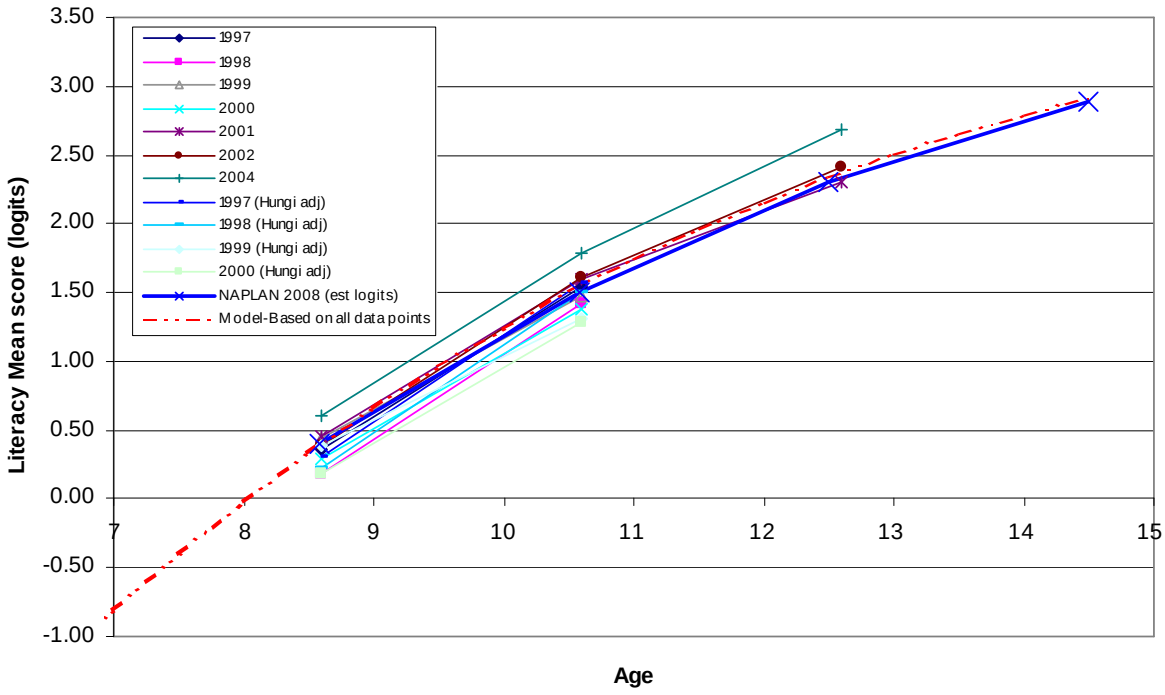
It is recognised that these assumptions are contestable. However a theory on what comprehensive test data might have looked like provides a comparison with the actual teacher data presented in detail in Chapters 7 and 8. Applying the above assumptions about how missing test data should look enables the development of the imputed data in the following sections. First frameworks for the group means for literacy and numeracy are established. Then typical cases are added for each missing Year level using combinations of actual students from the tested Year levels.

Setting the framework for the Literacy model

Figures 6.1 and 6.2 indicate two views of the data in Table 6.4. In both figures the X-axis is converted to age, with the Year level points plotted at the average age for the Year level. In Figure 6.1 the cross-sectional points within a calendar year are connected. On the whole their

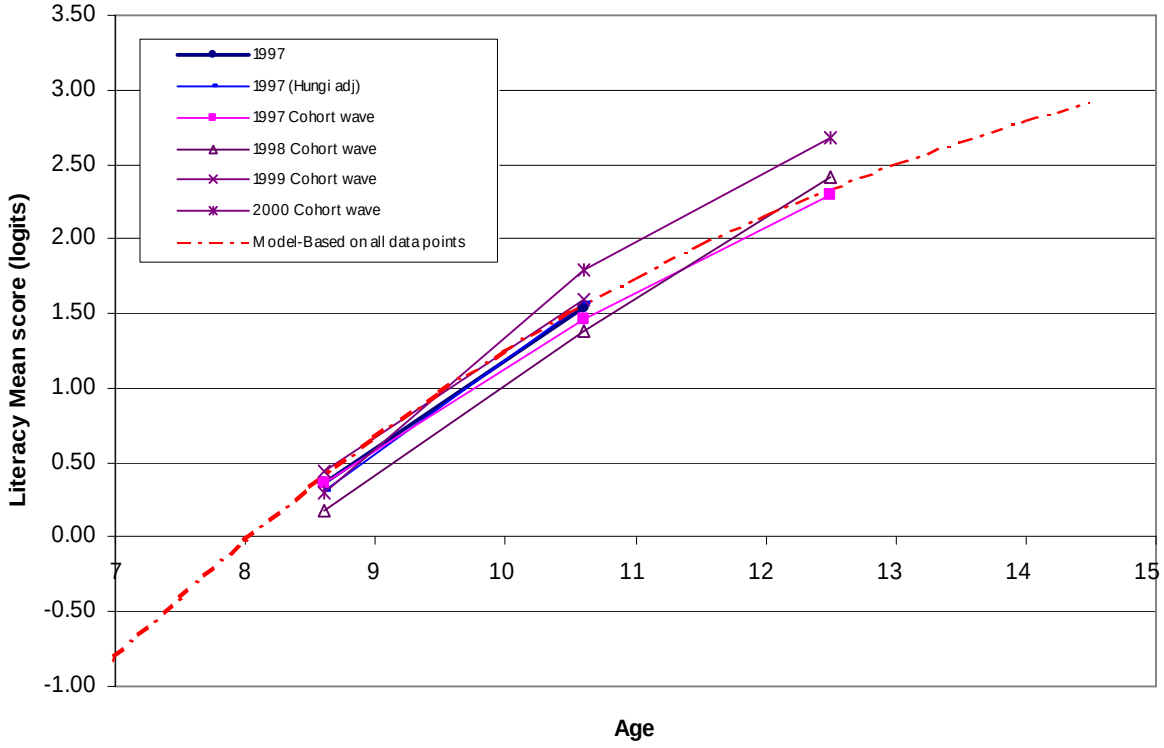
trajectories are similar even though displaced due to variability in the scaling and/or actual variation in the performance of the groups. The line for 2004 shows the same cross-sectional growth pattern as the rest but is displaced above the group. This is possibly related to a change in the testing and analysis provider and the resultant difficulty in aligning with the previous scale. The SA NAPLAN reading data (as distinct from Literacy) appear to follow the same general trajectory.

Figure 6.1 Literacy mean scores –Cross-sectional view with model trajectory



In Figure 6.2 the plotted lines link cohorts at two-year intervals. These follow less consistent trajectories. The greater variability in trajectory for these waves suggests that some variability is due to the variations in the scale with calendar year. Within calendar year linking appears more reliable than the across calendar year linking. Longitudinal views based on Hungi's re-scaled data follow the same general trajectories as cross-sectional data, suggesting that the variation in the longitudinal view of original scores is due less to variation in growth than variation in the scaling process.

Figure 6.2 Literacy mean scores –Longitudinal view with model trajectory



The dashed line is the trajectory obtained using the four points, including SA NAPLAN data, and fitting a Gompertz curve using CurveExpert (Hyams, 2001). To achieve this the raw logit values are scaled upwards by 10 logits to remove all negative values and to place the likely mean learning score value at age=0, very low. This is done, as the Gompertz expression cannot be fitted to negative values. Once a curve is fitted the values for each age point are then rescaled back to the original logit scale origin by subtracting 10 units[27]. Target means can then be identified for each Year level, using the fitted Gompertz curve. The target is calculated using the average age for the Year level. The derived curves are shown in Figures 6.1 and 6.2.

The next stage in the model development is to generate student records for each Year level using sampling of data points from the known data distributions for Years 3, 5 and 7.

Adding multiple points to Literacy model

Imputed data points for each Year level were developed in stages. The first stage required establishing the date of birth for as many records as possible. The original testing process did not collect date of birth, only conventional integer age at the point of testing. The lack of age

[27] The values of the parameters for the fitted Gompertz curve are a = 14.05, b = 0.65, c = 0.22. The estimated score value at any age value is converted to the original logit scale by subtracting 10. The asymptote (a) of the group mean is effectively 4.05 logits on the original scale as age moves above 20.

detail made an analysis of test data by actual age at testing impossible. To remedy this a date of birth was found for as many students as possible. This was a long process of matching names to a master file of names and obtaining the date of birth from the student file along with a unique student identification number. About two thirds of each Year level (8000 of 12000 records for each of four cohorts) were assigned a date of birth. The student identifier was required for later matching to teacher judgement assessments in Chapter 8.

The statistical characteristics of the original files and the subset that were assigned dates of birth are tabulated in Table 6.5. For Year 3 the mean of the sample with birth dates (8988 out of 12437: 72%) has a mean of 0.46 logits, 0.1 logit greater than the full cohort mean of 0.36. This indicates a slight bias in name matching against finding some lower scoring students. The standard deviation, inter quartile range, skewness and kurtosis values of the sample and the original cohort are similar enough to assume that they have similar distributions even though the mean is greater. In the model building process, the sample is set to a new mean by adjusting the value for each individual case by the amount required to make the model mean match the target mean from the framework.

For Year 5, 8651 out of 11972 records (72%) were matched. The mean for the sample with birth dates was almost identical to the full sample.

Table 6.5 Literacy-Comparison of original records with subsets assigned dates of birth

	Sample with Birth Dates		Full cohort	
	Year 3	**Year 5**	**Year 3**	**Year 5**
Mean	0.46	1.56	0.36	1.55
Median	0.51	1.69	0.41	1.69
SD	1.44	1.20	1.36	1.24
Skewness	-0.30	-0.70	-0.34	-0.79
Kurtosis	4.35	5.43	4.42	5.65
IQR	1.74	1.54	1.91	1.61
N	8988	8651	12437	11972
% with DOB			72.3%	72.3%

Once dates of birth were established, the ages of students were calculated and categorised for specific age categories relative to the middle of August 1997, the test period. Age was represented by actual age at testing, in categories of integer age (age last birthday), age in half years, age in 0.2 of a year and age in 0.1 of a year, approximating an age in years and months. Students were placed into the categories on the basis of the relationship of their actual age to interval boundaries. The 0.1 categories were centred on the required values with boundaries at 0.05 of a year. The 0.2 categories were centred on the odd values (0.9, 0.1, 0.3, 0.5 etc) with the even values being the interval boundaries.

Files for Years 3 and 5 were randomly sampled to select 7950 records for each Year level from the larger samples (8988 and 8651 respectively). This limit was required to ensure that the final model of records would fit into the version of Excel being used, which had an upper limit of 64,000 records, allowing a maximum of 8000 cases for each of Years 1 to 8. The mean of the Year 3 sample was adjusted slightly to match the target set in the framework model.

For the Year 4 component the original Years 3 and 5 cases with assigned dates of birth were sampled to select 3975 records from each source. These records were then summarised to obtain the initial Year 4 sample mean. These new data were then adjusted by the required amounts to set the grand mean to the framework target for Year 4. A common amount was added to the each of the Year 3 derived records and a common amount subtracted from each of the Year 5 derived records.

Year 7 data from 2001 and 2002, shown in Table 6.6, included dates of birth as part of the data collection and testing procedure. Cases from both 2001 and 2002 files were combined.

Table 6.6 Comparison of 2001, 2002 and 2004 Literacy score statistics - full cohorts

Statistics	2001 Y7 Literacy	2002 Yr 7 Literacy	2004 Yr 7 Literacy
Mean	2.30	2.41	2.68
Median	2.30	2.46	2.71
Skewness	-0.04	-0.26	-0.26
Kurtosis	3.40	3.54	3.45
SD	0.92	1.08	1.15
Inter Quartile Range (IQR)	1.17	1.43	1.48
SE (Mean)	0.01	0.01	0.01
N	12533	12069	15628

The grand means of two independent samples of the Year 7 composite records were set to the framework model targets by systematic individual record adjustment to generate records for Years 7 and 8. A problem was discovered later, well after the model data were developed. Students who missed one or other of the tests in 2001 were assigned the minimum possible score rather than deleted. This influenced the means. On discovery of the problem zero score records were deleted from the samples, leading to final samples being less than the intended 7950. Year 6 records were developed in the same manner as Year 4 records. Samples were drawn independently for half the required records from Year 5 and Year 7, and each subset adjusted to average to the overall framework required Year 6 mean.

The creation of records for Years 1 and 2 required one additional step. Both were based on independently sampled Year 3 subsets of 7950 records from the Year 3 records with dates of birth. The data were then summarised by 0.1 of an age and the notional mean scores for each

0.1 age category spread down the age scale to follow, in general terms, the steeper gradient at these age points in the model. The adjustments for Year 1 were greater than the adjustments for Year 2 and are justified on the basis of the age within grade/Year level observations described in Chapter 5. Evidence from Chapter 5 suggests that the within-Year level gradient with age does not match the general across age gradient. The within-Year level gradient gets flatter with increasing age but in the first years of school almost follows the general trajectory. The model records were adjusted to show this steeper gradient within Year level. This was done by setting the mean for each 0.1 age cohort to sit on the general trajectory line while keeping the mean for the full Year level cohort at the framework specified value. Records were adjusted manually and iteratively for each 0.1 age cohort leading to approximate matches only of the means to the target for each 0.1 age cohort. The matches of the Year level means at Years 1 and 2 to the framework overall are very close as shown in Table 6.7.

The effects of the score adjustments for each Year level in the model, relative to the target values to be achieved, are shown in Table 6.7. The general characteristics of the final 63,306 simulated students are illustrated. The framework target means and the means of the imputed points for each Year level match well. The inter-quartile range reduces with increasing Year level, as does the SD, as expected from Chapter 5. The SDs in Years 7 and 8 are effectively the same as the two samples are clones from the same Year 7 distributions.

Table 6.7 Literacy Model-main statistical characteristics

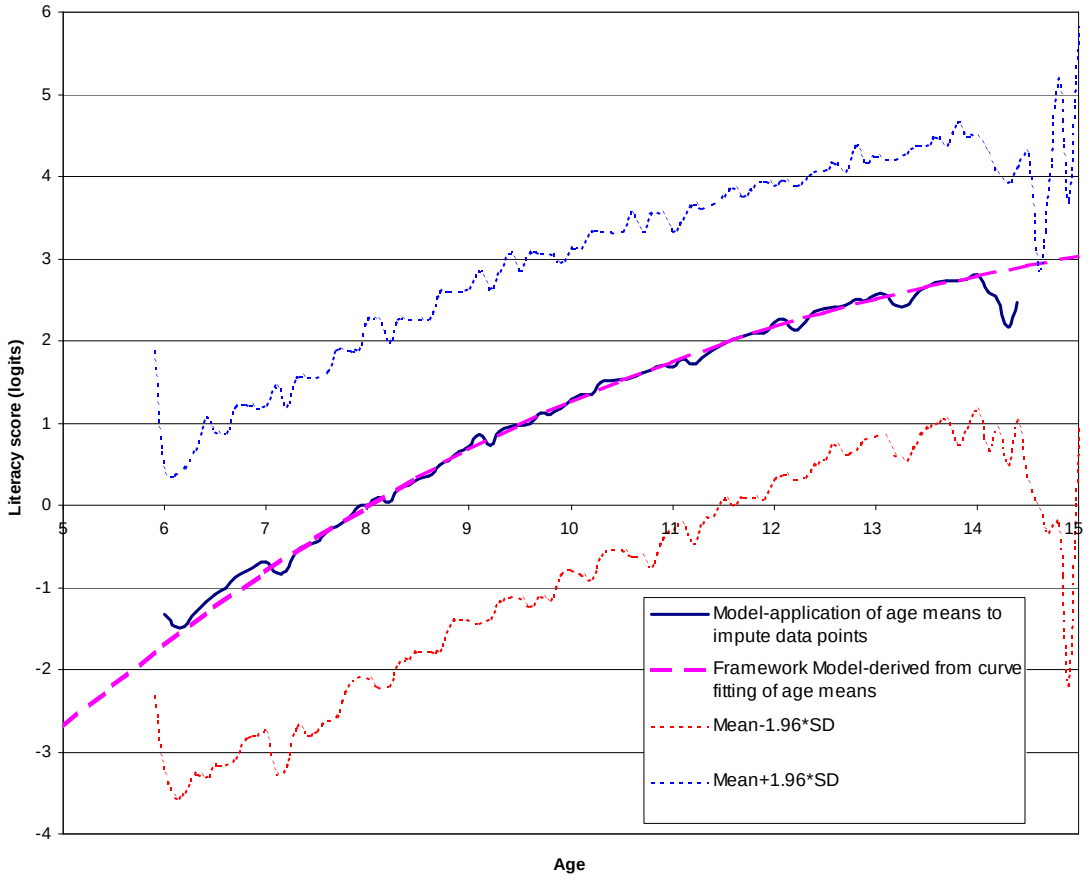
Literacy Model by Year Level	1	2	3	4	5	6	7	8
Framework (Targets for each YL)	-1.14	-0.31	*0.41**	1.04	*1.57*	2.01	2.39	2.69
Means of used or imputed points	-1.14	-0.32	*0.41*	1.04	*1.57*	2.02	2.39	2.70
Medians of used or imputed points	-1.10	-0.25	*0.47*	1.12	*1.69*	2.05	2.40	2.71
SDs of used or imputed points	1.35	1.32	*1.31*	1.26	*1.20*	1.07	1.00	1.02
Skewness of used or imputed points	-0.27	-0.32	*-0.29*	-0.45	*-0.63*	-0.26	-0.15	-0.20
Kurtosis of used or imputed points	4.12	4.42	*4.32*	4.70	*5.03*	3.67	3.63	3.77
IQR of used or imputed points	1.85	1.74	*1.74*	1.65	*1.54*	1.36	1.29	1.29
Count of used or imputed points	7949	7949	*7949*	7949	*7949*	7888	7837	7836

* Italics signify Year levels with actual data, although case values have been adjusted to average to the framework Year level means.

The Model data compared to the Framework

The complete data set for the imputed data points is summarised in Figure 6.3. The data are shown as if they are continuous but are discrete means at 0.1 of an age. The means of the scores at each age point, when calculated independently of Year level, follow a trajectory with age that matches the framework model. This is unsurprising at the lower Year levels since the data were adjusted to achieve this result. However from Year 3 onwards the trajectory of the means is determined by the natural elements of the data.

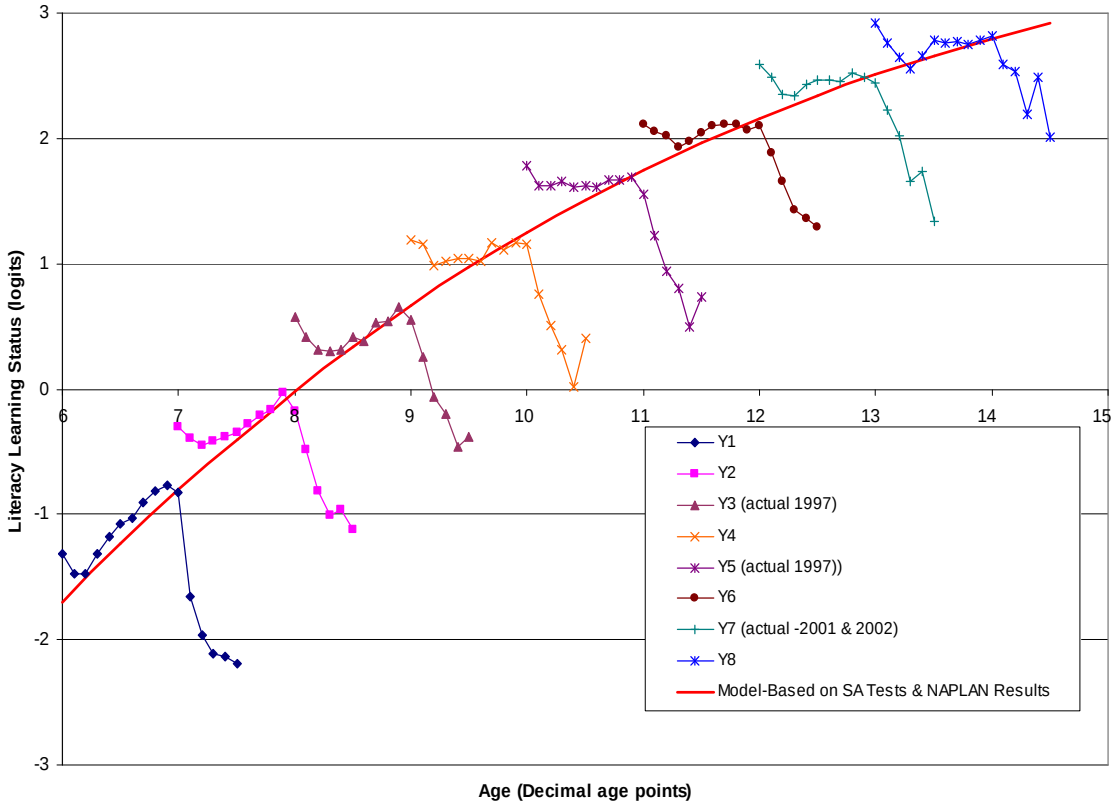
Figure 6.3 Comparison of Literacy Model to the Framework Model



The 2.5th and 97.5th percentile ranges are shown, to indicate that the general pattern with age applies across the spread of the data. Widening the age category (using 0.5 of an age category relative to 0.1) can smooth the fluctuations around the general trajectory but hides the elegance of the pattern with age.

When model data are analysed by age within Year level, the general trend of increased mean score with age within a Year level is revealed, as shown in Figure 6.4. Once again the trajectories at Years 1 and 2 are artificial. From Year 3 onwards these reflect the patterns in actual data. The effects at Years 4 and 6 are achieved by blending years above and below, which may not reflect actual data patterns.

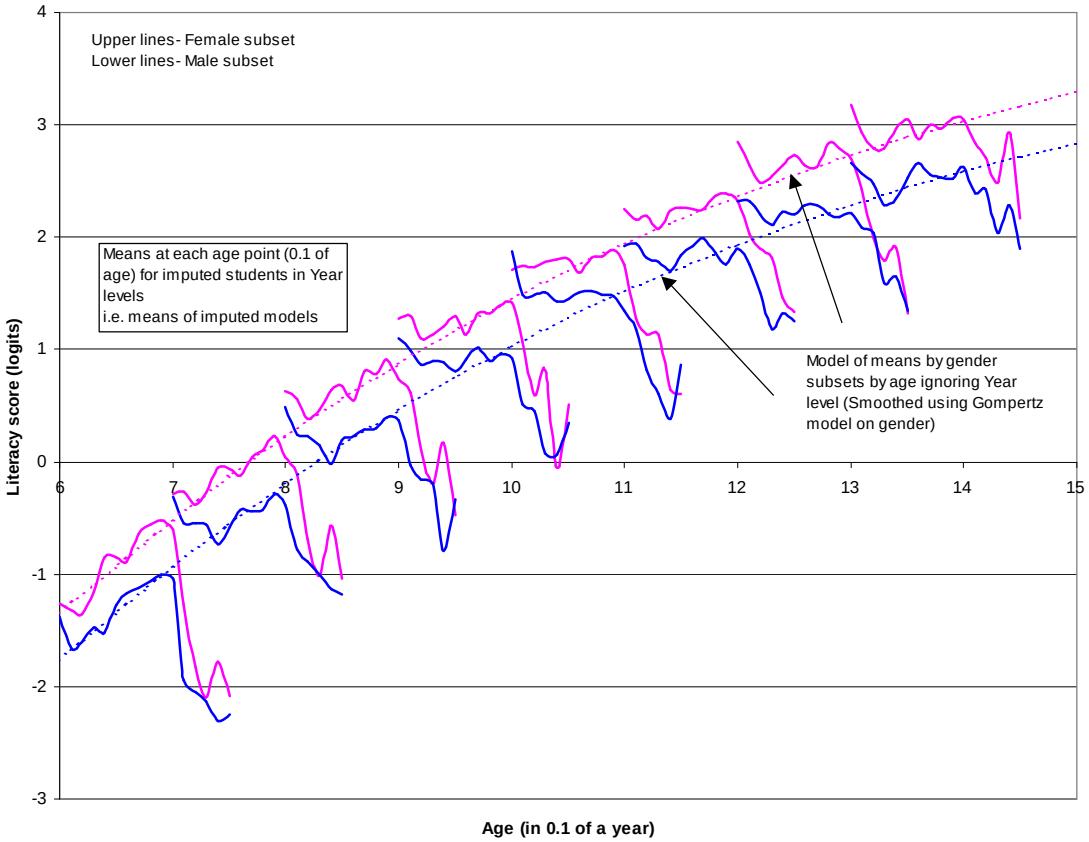
Figure 6.4 Literacy Model by Year level



The within Year level patterns in the model are censored in Figure 6.4: at both ends of a Year level, cases included in Figure 6.3 are censored in Figure 6.4. These censored cases are very small numbers of students with ages well outside the normal age range. The pattern by Year level is generally consistent. There is a very short lead-in for a very small group of younger children (not all shown) who appear to have higher mean scores than those only a month or so older. Then, for the bulk of students, the mean scores increase with age until the highest within normal age range age category for the Year level is reached. Then the set of older students, larger and covering a wider age range than the small group of very young students, produce a tail where mean scores drop off quickly.

Assuming the model approximates reality, it appears that the relative smoothness of the curve in Figure 6.3 is partly a result of the compensatory effects of the older and younger tails, and their proportionately small numbers in the data clusters for each Year level. If there was no age effect the Year level curves would be flat and the composite curve a more pronounced stepwise curve. The data model can be summarised by gender and Year level as shown in Figure 6.5. Unsurprisingly the gender effect is reflected in the model data.

Figure 6.5 Literacy Model by Year level and gender



In the Literacy test model the female mean performance by age is greater than the male mean performance and is consistently represented thus at each year level. The curves of the gender subsets follow the smoothed trajectories shown in Figure 6.5. The smoothed trajectory is obtained by applying a Gompertz model to the gender subsets independent of Year level. The model suggests that the gender effect is consistent over all Year levels, and increasing slightly with increasing Year level/age. The model shows a difference in favour of females is 0.41 logits at age 6, increasing to 0.44 logits by age 12.

The development of the model is based on expecting the mean scores for Year level cohorts to sit on an idealised trajectory. In the South Australian school system around 1997 it is speculated that the patterns identified in Figures 6.3, 6.4 and 6.5 approximate the data for a test assessment for Literacy, applied consistently from Year 1 to 8. This general speculation about Literacy is returned to once the companion story for Numeracy is developed and after the teacher-assessed view of the same learning development is described in Chapter 7.

The trajectory of Numeracy test scores

A similar process as applied for Literacy data is applied in the development of the model for Numeracy data from the 1998 Basic Skills test. The 1998 data are selected to match the

timing of the teacher assessments for Mathematics described in Chapter 7. The data considered in developing the framework are summarised in Table 6.8.

Table 6.8 Numeracy – Mean scores by Year level and Testing Year

	Year Level	3	5	7	9	Growth	Growth	Growth
	Average age	8.6	10.6	12.6	14.5	3 to 5	5 to 7	7 to 9
Cross-sectional	1997	0.05	1.30			1.25		
	1998	**0.13**	**1.34**			1.21		
	1999	0.18	1.27			1.09		
	2000	0.08	1.11			1.03		
	2001	0.13	1.18	2.28		1.05	1.10	
	2002	0.36	1.24	2.46		0.88	1.22	
	2004	0.61	1.46	2.53		0.85	1.07	
	Mean	0.22	1.27	2.42		1.05	1.13	
	1995 (Hungi adjusted)	0.30	1.21			0.91		
	1996 (Hungi adjusted)	0.31	1.24			0.93		
	1997 (Hungi adjusted)	0.21	1.31			1.10		
	1998 (Hungi adjusted)	**0.22**	**1.34**			1.12		
	1999 (Hungi adjusted)	0.20	1.36			1.16		
	2000 (Hungi adjusted)	0.15	1.24			1.09		
	Mean	0.23	1.28			1.05		
Longitudinal	1997 Cohort wave	0.05	1.27	2.28		1.22	1.01	
	1998 Cohort wave	0.13	1.11	2.46		0.98	1.35	
	1999 Cohort wave	0.18	1.18			1.00		
	2000 Cohort wave	0.08	1.24	2.53		1.16	1.29	
	Mean	0.11	1.20	2.42		1.09	1.22	
	1995 Cohort wave (Hungi)	0.30	1.31			1.01		
	1996 Cohort wave (Hungi)	0.31	1.34			1.03		
	1997 Cohort wave (Hungi)	0.21	1.36			1.15		
	1998 Cohort wave (Hungi)	**0.22**	**1.29**			1.07		
	Mean	0.26	1.33			1.07		
NAPLAN	NAPLAN 2008 Numeracy (estimated logits)	0.13	1.16	**2.24**	**2.74**	1.00	1.10	0.50
	NAPLAN SA 2008 Numeracy (reported scores)	*(388.8)*	*(460.4)*	*(536.2)*	*(571.1)*	*(71.6)*	*(75.8)*	*(34.9)*

* Bold values signify values used in estimation of Gompertz model parameters.

Setting the Framework for the Numeracy model

The 1998 views of Numeracy are shown in bold in Table 6.8. For the framework model growth values are the most critical. In the cross-sectional view the growth from Year 3 to 5 appears to have reduced since 1997. However the Hungi adjustment reduces the spread of the cross-sectional growth. Averaged over all views, the growth over two years from Year 3 to 5 is just over 1 logit, slightly less than the general growth for Literacy between these Year levels.

A key difference, relative to Literacy, is the growth in the next two-year period, Year 5 to Year 7. The growth rate for Literacy diminishes, while for Numeracy the growth rate is almost identical to that for Year 3 to 5, based on the cross-sectional, Hungi adjusted and NAPLAN data. Effectively growth in Numeracy learning is linear with age from Year 3 to

Year 7. Based on one point only (the NAPLAN data for SA), the growth rate in Numeracy appears to reduce from Year 7 to 9.

A curve is fitted to the points using a Gompertz iterated solution for the highlighted data for Years 3 to 9, at the average age for each Year level. The process is the same as for the Literacy framework model. Data points are increased in values by 10 logits to avoid any negative values, the Gompertz curve is fitted, and then the resultant curve is adjusted back to the original logit scale. The values of the parameters for the fitted Gompertz curve are a = 15.33, b = 0.37, c = 0.14. The estimated score value at any age value is converted to the original logit scale by subtracting 10. The asymptote (a) of the group mean is effectively 5.3 logits on the original scale as age moves above 20.

Figure 6.6 illustrates the wider spread of values in the cross-sectional view relative to the longitudinal view in Figure 6.7. The most different cross-sectional set is 2004, consistent with Literacy data. The 2004 tests and scaling were provided through a different contractor.

Figure 6.6 Numeracy mean scores-Cross-sectional view with model trajectory

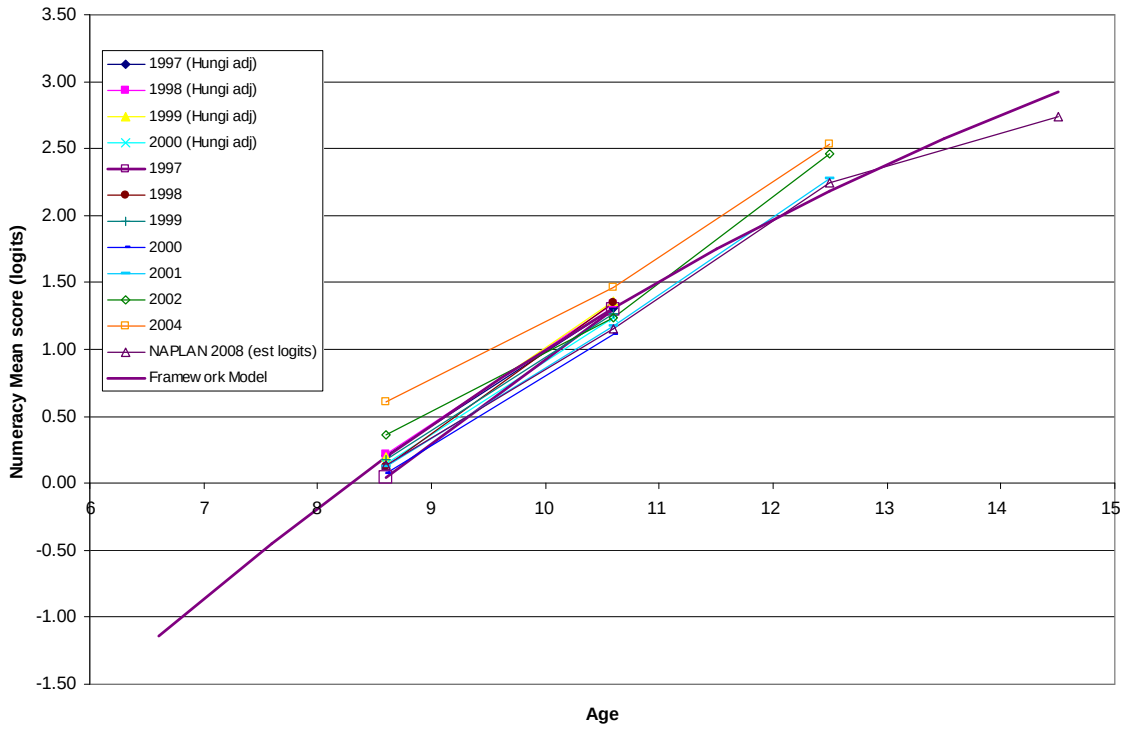


Figure 6.7 Numeracy mean scores –Longitudinal view with model trajectory

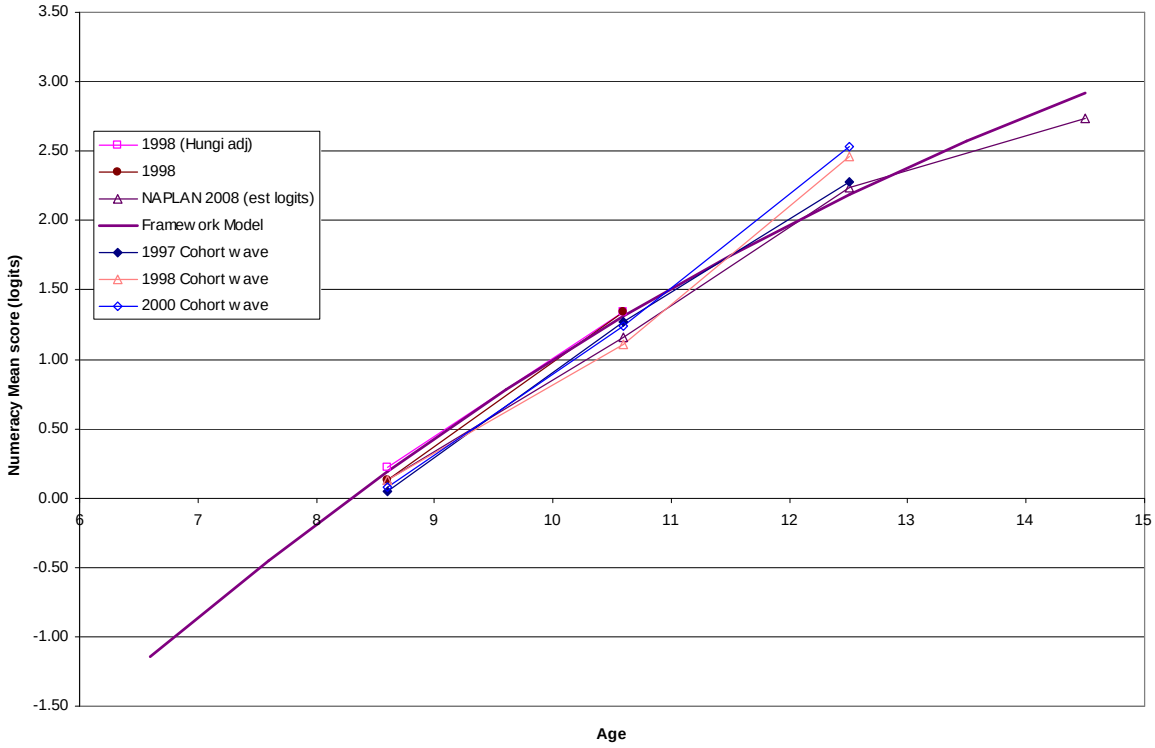


Figure 6.7 shows that the trajectories for most cases in this view are parallel. Consistent with the Literacy data, large differences occur in the means for the Year 7 data in 2002 relative to 2001. The model sets a framework for imputing points. The trajectory at Year 9 (age 14.5) for Numeracy may be poorly modelled. At Year 7 (age 12.6) the fit of the estimated NAPALN data to the trajectory is close.

Adding multiple points to the Numeracy model

As for Literacy, dates of birth were found through matching student records for as many cases as possible. Table 6.9 shows that there were slightly higher percentages of students for whom dates of birth were found than for the 1997 cases for Literacy. Almost 75% of Year 3 and 80% of Year 5 were assigned dates of birth. Ages were then calculated in the following age categories: actual age at testing, conventional age at testing, age categorised into 0.5, 0.2 and 0.1 categories of age at testing. The general statistical characteristics of the sample with birth dates are compared with the original 1998 cohorts in Table 6.9. Apart from the mean for Year 3, most characteristics are very similar.

Via the model framework a process identical to that for Literacy was used to add data points for students. The framework targets for the means of each Year level are shown in Table 6.10. The model was built from the 9567 and 10008 records for Years 3 and 5 respectively. Independent samples of 7990 were taken for Years 1 to 3 and 5. Original Year 3 and 5 cases

were sampled and then combined to create 7990 cases for Year 4. Year 7 files for 2001 and 2002, as used for Literacy, were used to create cases for Year 7, Year 8, and blended with Year 5 to create cases for Year 6.

Table 6.9 Numeracy-comparison of original records with subsets assigned dates of birth

	Sample with Birth Dates		Full cohort Statistics	
	Year 3	**Year 5**	**Year 3**	**Year 5**
Mean	0.20	1.38	0.13	1.34
Median	0.22	1.38	0.22	1.38
SD	1.32	1.12	1.36	1.16
Skewness	-0.09	-0.25	-0.15	-0.42
Kurtosis	4.39	5.85	4.54	6.38
IQR	1.71	1.35	1.71	1.35
N	9567	10008	12794	12471
% with DOB			74.78%	80.25%

A statistical summary of the match of the Numeracy model to the Framework model is documented in Table 6.10. The scores for each student were systematically adjusted to bring the mean for the year level as close as possible to the target. The spread characteristics reflect the original data sources. An anomaly in Year 7 data mentioned in footnote 25 also applied for Numeracy but for a separate set of students who had Literacy scores in the expected range but no Numeracy score. The anomaly was discovered after the model had been developed. As a result the cases omitted in the original files were deleted from the final model and the remaining records adjusted to match the Target means. This caused a small loss of records in the Year 6, 7 and 8 models, reflected in the count of points in these Year levels being less than the 7990 target.

Table 6.10 Numeracy Model-main statistical characteristics

Numeracy Model by Year Level	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Model (Targets for each YL)	-1.16	-0.45	*0.20*	0.79	*1.32*	1.80	2.19	2.57
Means of actual or imputed points	-1.16	-0.45	*0.19*	0.78	*1.31*	1.75	2.22	2.59
Medians of actual or imputed points	-1.16	-0.40	*0.20*	0.82	*1.32*	1.72	2.18	2.54
SDs of actual or imputed points	1.35	1.34	*1.32*	1.22	*1.12*	1.13	1.12	1.10
Skewness of actual or imputed points	-0.11	-0.06	*-0.03*	-0.01	*-0.28*	0.04	0.32	0.21
Kurtosis of actual or imputed points	4.34	4.38	*4.19*	4.51	*5.70*	4.93	4.10	3.80
IQR of actual or imputed points	1.74	1.66	*1.71*	1.51	*1.35*	1.37	1.44	1.44
Count of actual or imputed points	7989	7990	*7900*	7990	*7990*	7916	7851	7872

The model compared to the Frameworks- Numeracy

Figure 6.8 compares the model of individual student scores with the target framework. The model follows the target trajectory well. The path of the trajectory is similar to the Literacy equivalent (Figure 6.3). The target points sit on the curves of the means. The intermediate points wobble along the general trajectories as should be expected from the stochastic nature

of the learning process and the known within-Year level age patterns. As the points at Year 3 (age 8.6) and above are derived from actual data, the model is assumed to approximately match the distribution of scores that would apply if all students in the model had been tested. Below Year 3 the extent to which the model matches reality relates to the steepness of the trajectory in these Years. To achieve the match to the Gompertz determined trajectories, data points were spread more widely by age.

Figure 6.8 Comparison of Numeracy Model with the Framework Model

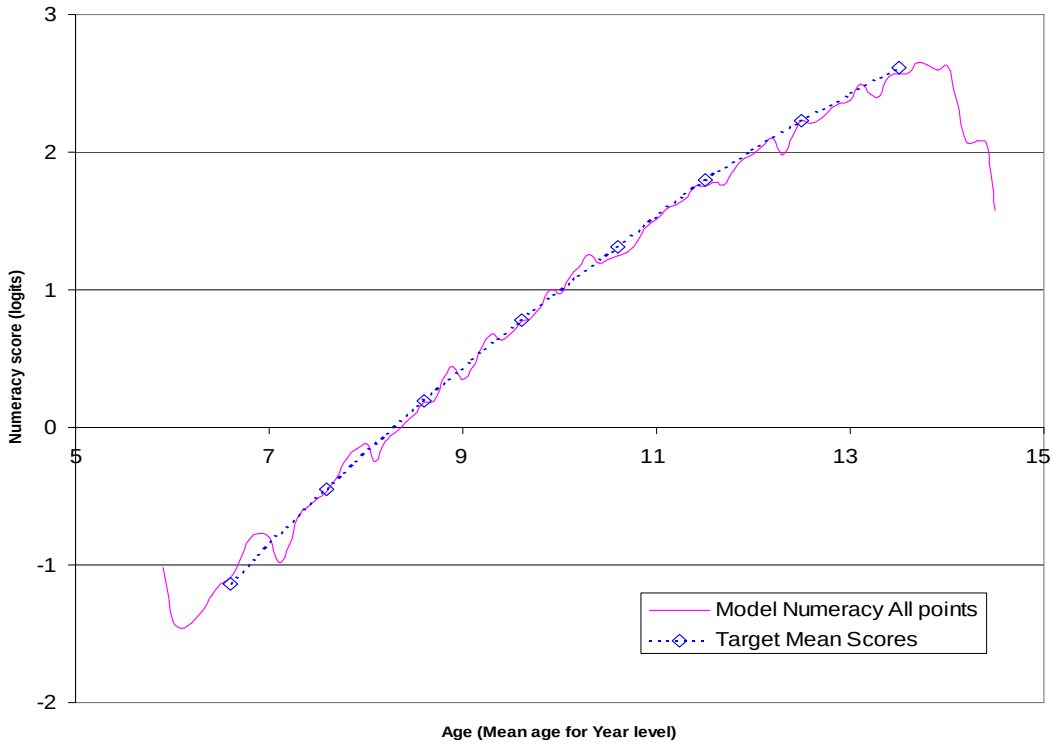
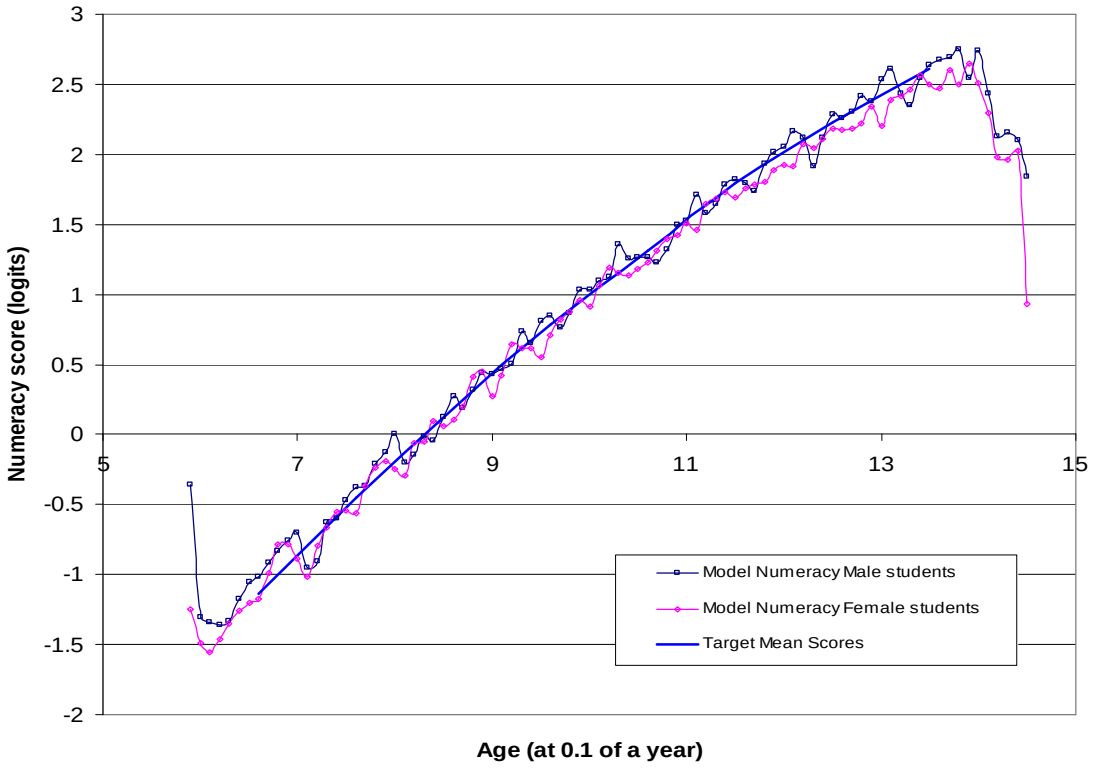


Figure 6.9 compares the trajectories of the mean points for each 0.1 of age, by gender. Recent evidence from NAPLAN (2008) indicates slightly higher scores for males apply, increasing with age (estimated to be 0.1 logits at Year 3 and 0.2 logits at Year 9). Data for 1998 indicate small differences of similar amounts (0.07 logits in favour of males at Year 3, 0.09 at Year 5). These differences contrast with the larger score advantage for females in Literacy at all levels, starting at about 0.3 logits and increasing with age.

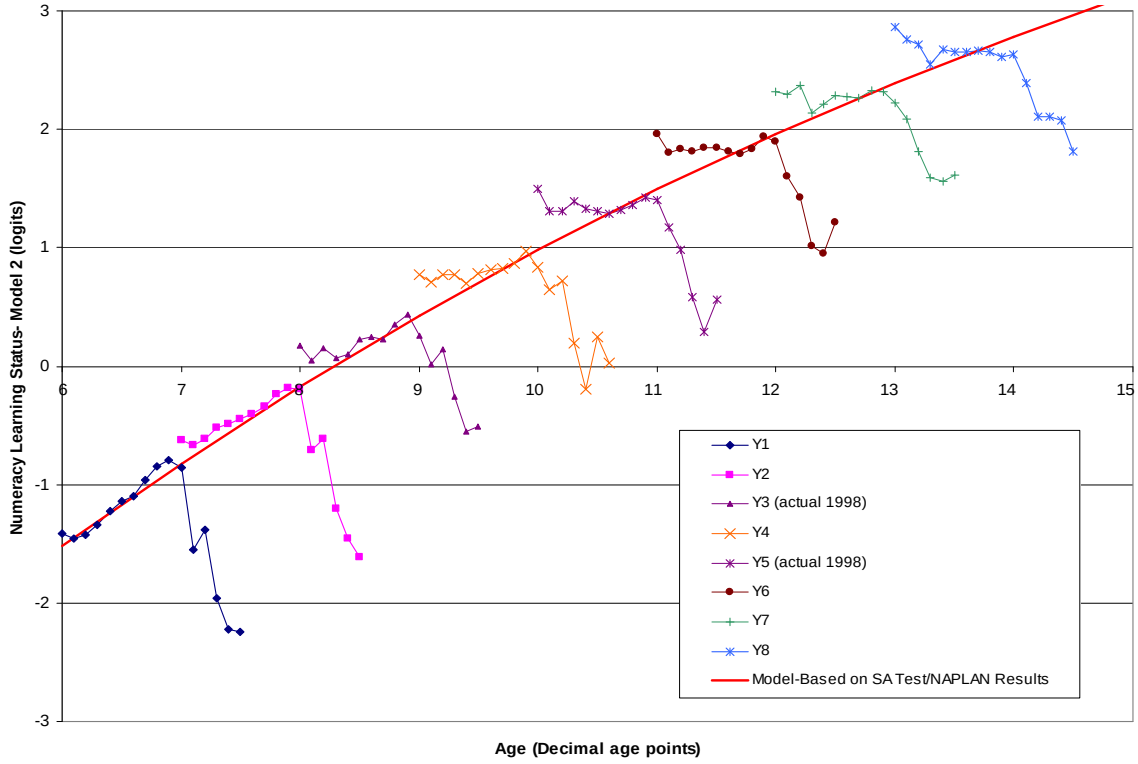
Consistent with the small advantage to males in Numeracy, the model generates a summary for males that tends, on average, to be greater than the female summary as shown in Figure 6.9. A more refined gender analysis from the model is covered in subsequent sections.

Figure 6.9 Numeracy Model by gender



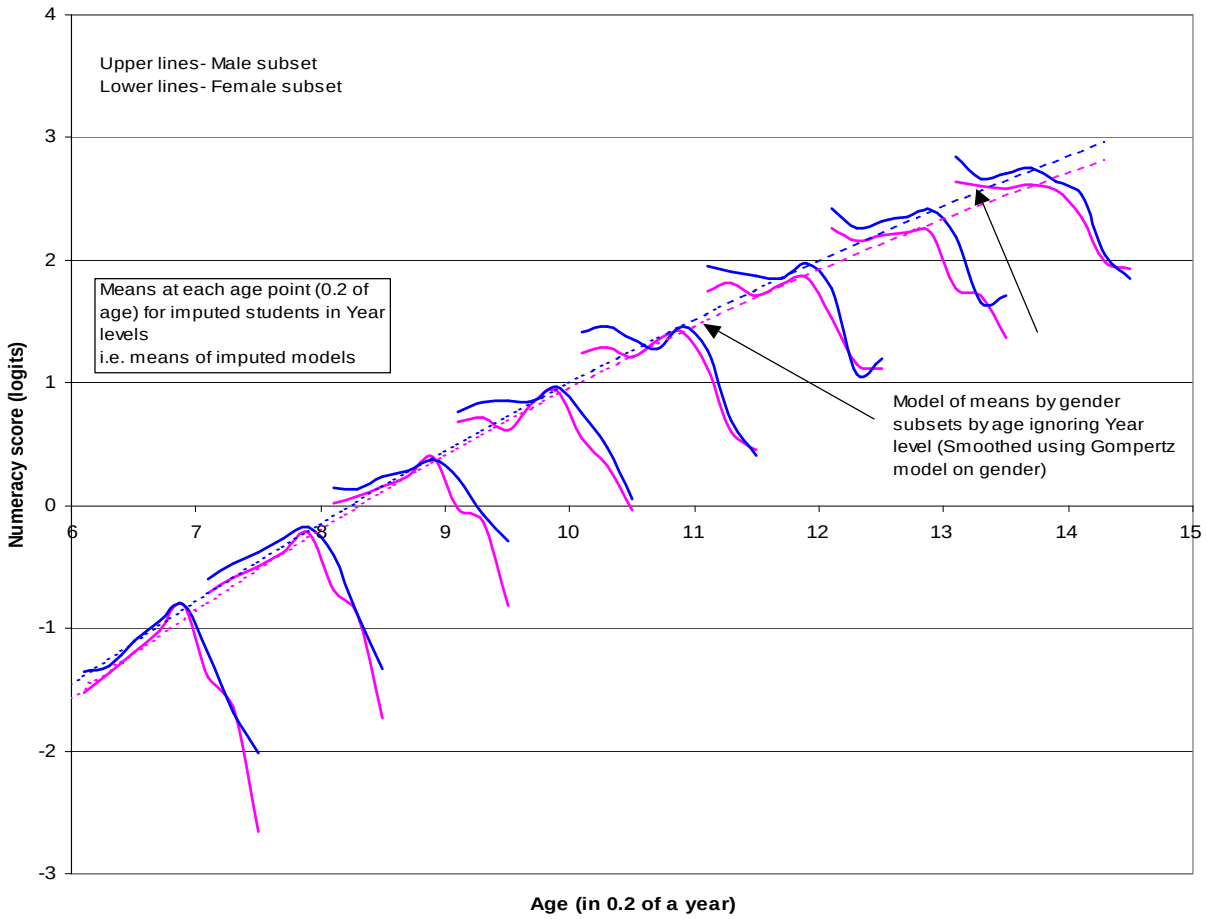
When the model is summarised by Year level (Figure 6.10) a very similar pattern is obtained for Numeracy as is found for Literacy. At lower Year levels the gradient with age within a year level appears to be greater (although for Years 1 and 2 the effect is artificially created by the model development process). At all Year levels, students older than the normal age range for the Year level have lower mean scores and generate a tail of diminishing scores. This is consistent with the examples reported in Chapter 5 where data summaries from a wide range of test samples show this specific pattern of learning status by age within a Year level. Also consistent with Chapter 5 the gradient of the effect diminishes with increasing Year level.

Figure 6.10 Numeracy Model by Year level



Year level data in the model can be disaggregated by gender as shown in Figure 6.11. The summary in this case uses age categories of 0.2 age divisions to smooth the variability shown in Figure 6.10. Consistent with the general understanding of performance by gender, the male performance in Figure 6.11 is marginally higher than that of females at each Year level, with the difference growing with age. The trajectory for each gender group can be obtained by fitting the Gompertz expression separately. The gender trajectories start close together with a slight male advantage as age (Year level) increases. This is consistent with the general pattern in the NAPLAN (2008) data.

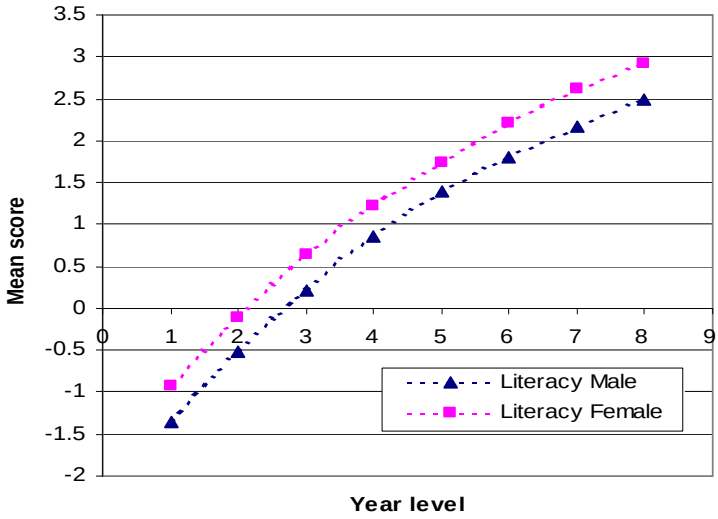
Figure 6.11 Numeracy Model by Year level and gender

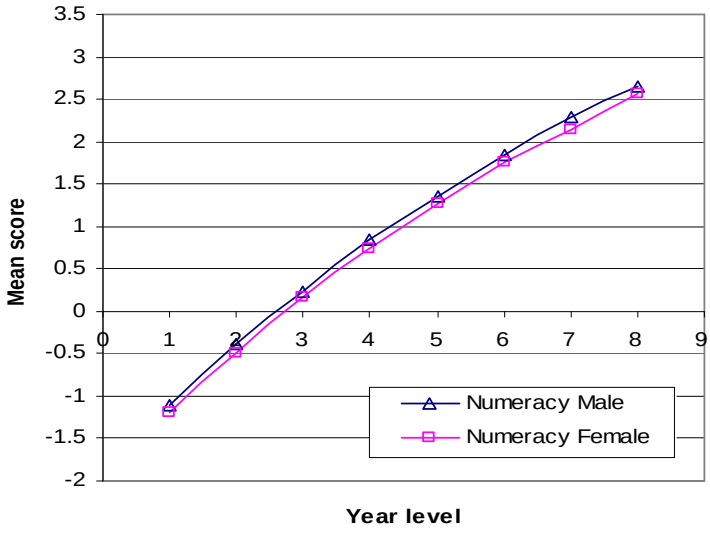


The models above were developed for two calendar years, for Literacy in 1997 and Numeracy in 1998, to estimate data for the non-tested Year levels and to understand the age effects within and across Year levels when consecutive cohorts of students are tested.

The panels below show Year level views illustrating the benefit of the models being developed on an individual student basis. Summaries can be made in a number of ways. In Figure 6.12 the Year level view is provided. The panels illustrate that the general models, developed without consideration of gender, enable a mean score to be estimated for gender by Year level or age.

Figure 6.12 Summary of the Literacy Model and Numeracy Model by Year level and gender





The gender views that come out of the models are consistent with those from other sources. Based on the UK Statistical First Release 19/2009 (2009, Table 5) for example, the trend by gender at Key Stage 2 (11 year olds) for mathematics shows the male average point score for 2009 at 27.7 points compared with 27.4 points for females (0.3 point difference), which has persisted for a number of years. On the other hand for English language there is a difference in favour of females of 1.6 points (28.1 for females, 26.5 for males). The specific patterns for both learning areas by gender have persisted for a number of years, at least since 2004 (Statistical First Release 19/2009, 2009). Accordingly, the models developed for the South Australian test data for Numeracy and Literacy follow, in general terms, the trends found elsewhere.

By building the model at a student level, a richer summary of the general patterns of performance by gender and age has been developed. For example the patterns by age within a

Year level can be shown. The models are based on an iterative fitting of a curve through four Year level score means (Years 3, 5, 7 and 9) using a Gompertz model. Other curve-fitting processes may produce equivalent results. The assumption of predictable growth in mean learning per Year level and by age is critical to the models. The evidence in Tables 6.4 and 6.8 indicates that average rates of learning growth with age have been approximately consistent across calendar years and where variations occur they can often be explained by test calibration variability. Assumptions about specific rates of growth with age and Year level would be unnecessary were assessment data available for all Year levels.

For two Year levels (3 and 5) the models use a large sample from the full cohort test data for the appropriate collection years. For Year 7 actual data are used, but are a blend of two collection years 5 years later. An estimated Year 9 mean influences the trajectory of the data.

Years 1 and 2 data points generated for the models are the most artificial since they are a transformation of individual scores from Year 3 to make cohort means match the framework. To achieve this a spreading of the data to increase the within Year level age gradient is required. The actual trajectory for the lower years is unknowable by the usual pencil and paper testing processes. The extrapolation from the known points, while plausible to the author at least, and generally consistent with trends from other sources as discussed in Chapter 5, is highly speculative. However the gradient of the lower trajectory is conservative relative to some estimates of the rates of learning at lower years (Hill, Bloom, Rebeck, Black & Lipsey, 2007 discussed in Chapter 5).

Summary

The purpose of the chapter was to report actual test data for Years 3 and 5 for South Australia and establish the quality of these test assessments. These data covering only two Year levels were then extended using the general findings for the trajectories of growth of learning status for cross-sectional groups (established to approximately match longitudinal groups) to develop a framework for the trajectory of the means at all Year levels. These framework trajectories were developed for literacy and numeracy.

Samples of student records taken from the actual Year 3 and Year 5 data for the appropriate calendar years (1997 and 1998), supplemented by Year 7 student records for 2001 and 2002, were then blended and means re-centred to fit the framework trajectories. Year 1 and Year 2 samples were stretched to match the framework trajectories and to match the general shape required by the models developed in Chapter 5.

The general data developed were then summarised from age, Year level and gender perspectives to report an estimated but speculative view of what summaries of learning status

of a sample of students tested at all Year levels form 1 to 8 might look like. These summaries provide one basis for comparing teacher judgement assessments of the same cohorts to test assessments.

In the next chapter the same general learning areas are assessed but based on teacher judgement assessments rather than tests. How the two approaches compare is addressed in Chapter 8.