Chapter 5: The trajectories of learning, growth and growth indicators

> Whatever relates to extent and quantity may be represented by geometrical figures. Statistical projections which speak to the senses without fatiguing the mind, possess the advantage of fixing the attention on a great number of important facts.
>
> Alexander Von Humboldt, 1811.

> Most of the quantitative research in educational psychology has been concerned with the microscopic processing of items by students or with the characteristics of tests. Without doubt, much has been accomplished in both of these areas- the first in terms of learning theory and the second in terms of test theory. What has been missing is a theory of a student's broad progress through a given curriculum.
>
> Suppes, Fletcher & Zanotti, 1976, p. 126.

The main purpose of this chapter is to describe the trajectories of learning status as Year level or age increase, using test data, i.e., to understand the general pattern of growth of learning over time. As a result, along with a deeper appreciation of what test data has added to our knowledge of longitudinal learning trends, the trajectories enable the development of frameworks for models of learning status data. These frameworks are then used in Chapter 6 to impute test data for ages and Year levels not actually tested in South Australia in 1997 and 1998. The sets of actual and imputed data allow a clearer basis for comparisons with teacher judgement assessments in Chapter 8.

Cross sectional and longitudinal data from Australia, US and England are compared to investigate whether trajectories of learning have common characteristics. These data are also explored as part of the question from Chapter 1: "What if teacher judgement assessments could provide the critical data needed to optimise the learning growth of every student?" If teachers had data for each student at a number of points throughout the school year, and access to the complete history of a student's previous data, what might it look like? How might a teacher make sense of such data?

There are very specific patterns in the mean learning status by decimal age (approximately equivalent to age in months) within Year levels that are revealed through test data. These patterns are described as a basis for comparing them with teacher judgment assessment data. The proposition is that if teacher judgement assessments are directly comparable to test data, then the same patterns of mean learning status by decimal age within each Year level should be evident.

The rates of learning as well as the path taken for each student are attributes for which teachers need a context. That context might be found in the Fullan et al. type of knowledge base (Fullan et al., 2006, p. 82) introduced in Chapter 1, using contributions of the sort identified in this chapter along with teachers' observations and other assessment data.

The chapter considers briefly the differences between the patterns of learning growth displayed by (the means of) groups of students compared with the much more variable patterns of individual growth. Because it has been logistically difficult and expensive to develop longitudinal learning status data for individual students through standardised processes, the general patterns of individual growth over short time intervals are not well appreciated.

Were teachers to generate individual student data, a further problem could be anticipated. Where will the research data needed to support teachers' understanding and diagnosis of unusual trajectories come from? How could teachers discriminate between expected variation and stalled trajectories? What action should they take when they do? What strategies are known to be effective? The knowledge base raised above would be an online source for teachers to explore these concerns.

The chapter concludes with brief references to two examples using analyses of test data as one source of contributions to the knowledge base on learning pathways. The examples establish learning maps of the likely order of learning numerals and letters. They are small but pertinent examples of ways in which test data could help provide reference frames for recording learning status progress as part of the support required for teachers.

Establishing trajectories of learning growth with age and Year level

Classroom assessment scoring systems do not place assessments on a vertical scale. This means that records from classrooms over extended time periods do not provide an adequate basis for an understanding of the variability of learning growth for individual students, or the general trend for the class. Currently there are very limited data sources from which to build and develop the insights and support for teachers that are required, were teachers to have the opportunity to put longitudinal records on a vertical scale.

Times series and cross-sectional data from statistical collections and general investigations using vertically scaled data, offer the beginning of an understanding of the pattern of learning development with time, or its proxies, age, Year level, and cumulative years of schooling. These data establish a basis for estimating (imputing) missing test scores as required in Chapter 6.

Test data investigated in Chapter 6 are available for Years 3 and 5 only in the calendar years of the teacher assessment collections. The actual and then imputed test score data developed in Chapter 6, are used to provide a comparison to the much richer estimates of teacher judged developmental position for students using levels scales for all Year levels from 1 to 8. As indicated earlier the trajectories are also important as elements of a knowledge base for teachers.

Considering the growth in learning status with time- the vertical scale

Before dealing with the time dimension (the horizontal axis), the scale of learning (the vertical axis) is considered. For the graphical and mathematical representation of learning with time without distortion, the units on both scales should to be equal interval. A necessary condition for a valid graphical representation of learning status over more than one Year level, or over time generally, is the

> construction of an outcome variable … measured on the same scale at every age of interest. Without such an invariant outcome metric, discussion of quantitative change or growth is meaningless. Standardized tests often fail to provide such a metric; different forms of a test are constructed with age-appropriate items for different age groups, and no effort is made to equate the forms. However, by calibrating the items across alternate forms, it is possible to construct a common measure for studies of cognitive growth. (Raudenbush, 2001, p. 508)

This linking of scales can be achieved (in principle) in the Rasch model (sometimes called a one parameter IRT model) through the use of common items as links to the scale properties of adjacent segments of the scale, overlapped to extend the scale vertically (Lee, 2003; Masters & Mossenson, 1983; Patz, 2007; Wright, 1977). The same principle is used to align scales when linking parallel tests (Hungi, 2003). The calibration can be arranged for adjacent Year levels/grades or test levels where tests have gradients in average difficulty within and across grade levels, or more comprehensively with a concurrent calibration of all tests in the one model-fitting process (Wright, 1997). The linking can be established through common items or through common students, that is the same students taking two or more forms of the test.

The adequacy of linking processes is contested. Haertel (1991), Holmes (1982) and Slinde and Linn (1978) present evidence and arguments that the linking process, using IRT scales, may not generate equal interval scales across the extended scale. Gustafsson (1979) challenges the Slinde and Linn (1978) analysis using a simulation study and concludes that the poor result in "vertical equating may be due to the fact that their treatment of the data introduced a poor fit to the Rasch model" (Gustafsson, 1979, p. 156). Gustafsson found that the Rasch model could be used to produce an adequate vertical scale as long as there is no correlation between item discrimination and difficulty. Similarly Guskey (1981) found that Rasch item calibrations and the Rasch ability scale were consistent and stable across test

levels. These results, he proposed, provide a model for the cross-validation of other test-scaling and score-equating procedures.

Stability of item difficulty over time is another critical element of the feasibility of a vertical scale. For a scale to remain consistent over time requires that item difficulties remain stable. Kingsbury (2003) has established that items can retain their relative difficulties as remarkably stable over periods of up to 20 years. In a similar fashion Griffin and Callingham (2006) have established the stability of a mathematics construct, tested with 14 year olds also over a 20-year span.

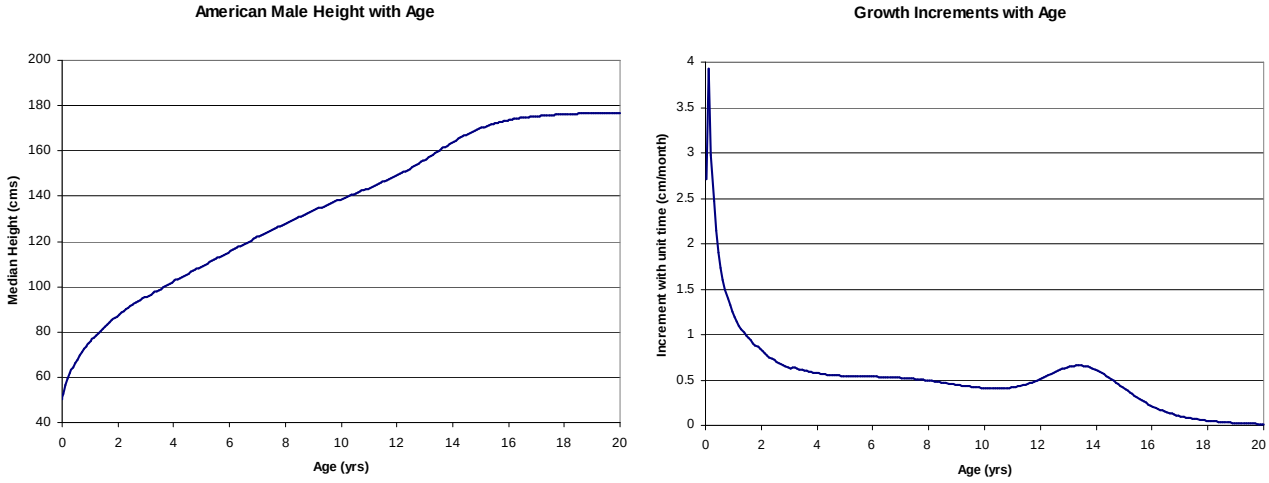
Scale shrinkage, an observed reduction in variance from early in the school year to later in the year and as the Year level increases in some tests (Yen, 1985, 1986), is also seen as a risk to vertical scales (Camilli, 1999; Camilli, Yamamoto & Wang, 1993). For some other tests e.g. the Iowa Tests of Basic Skills, the variance increases (Hieronymus & Hoover, 1986; Petersen, Kolen, & Hoover, 1989). The contradictory variance patterns are seen by Camilli (1999) and Camilli et al. (1993) as a challenge to the adequacy of the derivation of the vertical scale.

Schulz and Nicewander (1997) however report that the "discrepancies noted between variance trends in grade equivalent and IRT metrics … are exactly what one would expect if the θ [the difficulty/learning metric] growth rate of the norm group for the tests used in these studies were negatively accelerated." (p. 329). Grade equivalent scales demonstrate linear growth and increasing variance with age/Year level. On the other hand most IRT derived scales, as will be illustrated later in the chapter, appear to demonstrate negative acceleration for normed groups. Schulz and Nicewander provide an elegant example of the link of growth spurts, in the occurrence of these growth spurts in height for young males and females at puberty. They demonstrate that the variance in height increases during the growth spurt period and then reduces as age increases.

While physical growth provides a salutary example of normed growth patterns, the shapes of the two curves (physical height and learning) differ in subtle ways. The pattern for height growth however serves as a useful contrast reference frame for growth principles. In Figure 5.1 the left panel illustrates the general trajectory of height with age and the right panel the rate of height growth with age, derived by subtracting each n^{th} point from the $n^{th}+1$ point (Center for Disease Control, 2000). The rate of height growth is high in the first month of life and then decelerates from that point until puberty where the rate accelerates for a brief period. As will be illustrated in later examples, learning growth curves exhibit a strong similarity to sections of the height curve. Apart from the two growth spurts (first month and around age 11.5) the height growth rate is decelerating for the majority of the time.

An asymptote in height is reached at about age 17-18 (for males). In spite of continuing energy intake, height remains at the asymptote for the individual; excess energy intake is converted to body mass rather than height. Growth in learning, measured on the Rasch approach of log odds of item difficulty, appears to have an approximately similar tendency to an asymptote for skills such as reading (and possibly numeracy skills). This plateau may be a ceiling effect due to the limited upper spread of difficult items but most tests cited later have provided an appropriate range of difficult items. The trajectories are illustrated in later sections of the chapter where models based on National Assessment Program–Literacy and Numeracy (NAPLAN) and US data sets are developed. . This of course does not imply that learning is genetically determined but does suggest maturational effects in learning rates. The purpose of the analogy is to illustrate that growth rates, usually consistently reducing with time and then with spurts (first months and puberty) have parallels with the general trajectory of population summaries of learning over time.

Figure 5.1 Physical Growth Curve of American Males-median curve.



The units for the measurement of learning

In most of the data referenced in this chapter the measure of learning has the logit as the basic unit. The length of the logit unit is consistent across the whole scale but the logit (unlike a centimetre) is not universally fixed when new bench marking events are added to the scale[20] (Linacre & Wright, 1989). The comparison of measures from tests intended to measure performance on the same scales requires "not only adjustment for differences in local origin, but also for variation in the substantive length of the measurement unit we have constructed for the underlying variable" (Linacre & Wright, 1989, p. 55).

[20] In principle it should be possible to define a standard for a logit (however transformed) in say reading comprehension, and convert all other appropriately developed scales to this standard logit. The Lexile (Stenner & Stone, 2004) could be regarded as an example of a 'standardised' unit.

The conventional process for publicly reporting data measured on a logit scale (accepted as consistent in length in a particular test and scale development) is to transform the basic logit unit of the scale into a new form. The transformation usually increases the numerical representation given to points on the scale to three digit numbers and places the lowest scale position so that the scale only takes positive values. The transformations do not alter the relative distances of the points on the scale and thus there is no impact of linear transformations of the logit-based scale on the general shape of learning development with time (Bond & Fox, 2007, pp. 206-217). Under these adjustments the generic shape of learning with time for a particular population, that is the relationship of the measure for normed groups with Year level, remains effectively the same even though the parameters of the shape will vary with the transformation. This holds as long as the transformation is linear.

Research evidence for unidimensionality across Year levels

The vertical scales of the Literacy and Numeracy tests used in Chapter 6 are shown to be appropriate by Hungi (2003). The concept of unidimensionality is key to test scales being able to be vertically scaled. Fulfilment of the requirement of unidimensionality is a matter of degree (Bond & Fox, 2007, p.140). Test data drawn on later in the chapter, to develop models of learning growth with time, have been derived from tests where some evidence is provided that the constructs can be scaled vertically.

For a construct to be vertically scaled it is required to be unidimensional. As the difficulties of items deemed relevant in the development of a vertical scale are increased, the unidimensionality however might be compromised. A number of procedures for establishing the likelihood of unidimensionality have been adopted. These include factor analysis, marginal maximum likelihood, covariance structure analysis and local item independence using Yen's (1984) Q_3 statistic (Alagumalai, Keeves & Hungi, 1996; McCall, 2006). Under the Yen approach, when the effects of the latent trait are taken into account, the correlation of the residuals of response pairs should be zero. In this case the unidimensionality requirement is satisfied and responses exhibit local independence (Yen, 1984). Alagumalai et al. argue the benefit of linear structural equation analysis and the use of disattenuated correlation. Stenner (1996) applied disattenuated correlation in supporting unidimensionality for reading.

Wang and Jiao (2009) acknowledge that vertical scales are widely used to measure students' achievement growth across several grade levels and, as described above, have been considered as having disputed psychometric properties. Particularly disputed are unidimensionality and construct equivalence across grades. They claim their work is the "first study to investigate invariance of construct of vertical scale using real data" (Wang & Jiao, 2009, p. 773).

Wang and Jiao investigated the factorial structure for each grade and the equivalence of the factorial structure across grades using data from the Stanford 10 National Research Program. Data were available for all grades from 3 to 10, with 1700 to 3200 students per grade. Using confirmatory factor analysis (CFA), the assumptions of measurement invariance and construct equivalence across grades (using structural equation modelling with AMOS) were studied to determine the adequacy of fit to a one-factor model. The one-factor model had previously been asserted (Wang, Jiao, Brooks, & Young, 2004) to have the best statistical and psychometric characteristics relative to other models.

Wang and Jiao found that the vertical construct of the Stanford 10 test is unidimensional for each grade and across grades. There was no construct shift across grades and the common construct of the test was the same construct across grades. For this vertical scale, at least, the possibility of applying a common scale across Year levels appears feasible.

The potential for reading to be considered as a unidimensional construct over an extended scale is supported by findings that

> Evidence seems overwhelming that we can usefully treat reading ability, readability, and comprehension as if they are unidimensional constructs. The strongest support for such a treatment comes from the fact that when reading data simultaneously fit the Lexile Theory and the Rasch Model, then differences between two reader measures can be traded off for an equivalent difference in two text measures to hold comprehension constant. (Stenner & Stone, 2004, p. 33)

While there is strong evidence that vertical scales can be valid in principle and that some scales can be assumed to be valid in practice, there are also critics of the calibration of vertical scales as described earlier. For the purpose of the sections that follow, the author has assumed that it is reasonable to accept that the examples can be considered to have equal interval properties. Without an assumption of approximate unidimensionality the concept of a vertical scale cannot be considered for more than small segments of the age/Year level spectrum. The test designs have included linking items to vertically link the scale segments in all cases.

The purpose of this scene setting for the vertical axis is to establish that, in principle, appropriately developed learning scales can be assumed to have equal-interval properties. As a result they can be used in the development of models for generalising the shape of average learning development on a Rasch model developed vertical scale over an extended time period (10-12 years). The trajectories of the mean learning status of Year level cohorts as Year level increases, obtained from the use of the vertical scale, provide a general understanding of some dynamics of learning.

Growth in learning status -the time (horizontal) dimension

The second dimension, time, is also considered before the plots of trajectories are developed.

Time is treated in a number of ways in time-related analyses or descriptions of learning. It can be considered as a continuous variable or as a category or as a level in a multilevel analytic model. It can be represented in the usual units of time (minutes, hours, days, months, years) with data points positioned directly on the time scale, or compressed and centred to categories such as Year level, integer age rounded down (i.e. age last birthday), age in months or decimal age (age represented as years and part years in decimal form) as examples. Time can be transformed to a log scale to make the log time logit relationship linear (Lee, 1993; Rasch quoted in Olsen, 2003, pp 61-70), through a 'metameter' (Rao, 1958). The time dimension can be treated as intervals of equal learning, 'isochrons' (Courtis, 1929, p. 690).[21]

In the most conventional model of measuring learning over multiple time points (in either a longitudinal or cross-sectional design), the time dimension is usually centred on a point for each test event. As a result the points represents Year levels, waves or group mean ages. All students for a 'time point' are tested at the one time (give or take a few days) with the set of test events spaced on the X-axis by Year level, mean age or the elapsed time between waves. Where rates of change are considered as part of the analysis, the rates can be calculated only where the time dimension is represented in an equal interval form. The centreing process for categories may influence the rate estimate if it biases or distorts the time metric. Estimating the relationship of learning to time or age requires a number of points on the X-axis.

Year level (or Grade level), as indicated above, is one option for the X-axis scale. This scale has equal intervals, assuming testing at the same time for each Year level, since the unit is effectively calendar years. However when data for groups of students (and for individual students) are plotted against the vertical scale of learning, the scale obliges the origin to be placed at an inappropriate point, based on the initial Year level of the school system.

---

[21] Rogosa & Willett (1985) acknowledge Rao (1958) as the source of a transformation of time ($1-e^{-\gamma t}$) to 'linearise' the curve, and describe the transformation as a 'metameter' of time. Rao in turn acknowledges Rasch (Rao, 1958, p. 3.; see also Olsen, 2003, p.61-70) as the source of the idea for the 'metameter' for time. Independently, Courtis who advocated the Gompertz curve (Johanningmeier & Richards, 2008, p. 236, also Chapter 3 this thesis) as a model for individual growth with time, developed an alternative time approach, the concept of the 'isochron' whereby the distance in a learning curve from start to asymptote 'could be divided into one hundred equal units'. Using the Gompertz equation as the basis, the inflection point is at 1/e, 36.79%. Growth from 0 to 10% is in a period of 10 isochrons, 10% to 48%, another 10 isochrons, 48% to 80%, another 10 (based on Johanningmeier & Richards, 2008, p. 236). By 50 isochrons the growth is at just above 97%.

Singer and Willett (2003) consider the options for time dimension and advocate the use of "sensible" units of time (p. 140). Singer and Willett discuss the options of wave (equivalent to Year level), actual age and age group in the context of multilevel models. They advocate the more useful variable age, "because it provides more precise information about the child at the moment of testing" (p. 140). Where this is feasible this thesis applies the same convention.

To provide a more appropriate time origin (from the author's perspective) data points at Year levels are plotted, in most cases, as notional ages, through converting the Year level to age. One process to do this is to plot points at the mean age for the Year level cohort at the point of testing. This transformation implies an origin of 0 age, at the notional point of birth. Curve fitting and models of growth through mean or median points for particular time points, have a greater face validity through this origin than through one that places a time origin at, or one time unit before, the initial category. The latter convention applies when Year level or wave approaches are used. Based on the form of curve fitted the vertical scale has an extrapolated value at the time origin. In some transformations of the test score to a scale this value at time 0 can be assumed to be close to zero learning, but the position of the true Y-axis zero (no learning) remains problematic.

Lee (1993) has rather speculatively projected reading and mathematics scores by age, to estimate the status in each of these constructs at zero age. To make the trend of learning linear with age, Lee rescaled the age dimension as log (age+1 year) for reading and log (age+2 years) for mathematics, akin to the Rasch metameter time transformation (Olsen, 2003). Lee concludes that the absolute zero of reading is at birth and at conception for mathematics. The Lee model presumes a linear relationship of the learning status with a log transformation of time. However actual data points exist for ages from 6 to 14 only. The extrapolation to age 0 might take a different pathway to that speculated by Lee. The issue of possible learning trajectories from age 0 to 15 is taken up again later in this chapter.

A note on cross sectional versus longitudinal data for trajectory of learning models

Data sources examined for indications of the trajectory of learning on a logit unit item difficulty scale fall into two major designs. The more readily obtained data are from cross-sectional surveys regularly applied by some school systems (Australia for Year levels 3, 5, 7, and 9; US States for Grades 3 to 8 as examples).

A smaller number of projects and collections take a longitudinal approach. These designs track the members of the same cohorts through successive assessments (annual or biennial) providing a longitudinal perspective. Given the general stability of cross sectional means over time the cross sectional patterns are assumed to approximate the longitudinal patterns.

The longitudinal cases should provide a more reliable illustration of the general trajectory of learning with time. However Hilton and Patrick (1970) established that the general change from testing period to testing period was similar whether the group of interest was cross sectional, longitudinal-not-matched (the cohort not adjusted for losses and gains), or longitudinal-matched (the members of the group included only if they have data for each test period) at lower grade levels. However at higher grades, as the impact of dropouts affect the cohort, the apparent growth rate in learning is inflated through the loss of, most often, the less well performing students. In the cases explored here most cohorts remain intact up to Year level/Grade 8.

Learning growth in cohorts - examples of growth trajectories for the test score means of groups of students

The next section of the chapter explores data from three countries; Australia, the United States and England. Each provides evidence for a curvilinear relationship of learning growth with Year level and age, as distinct from a simple straight-line model of growth with time. These examples illustrate the general relationship of mean learning status for Year level or grade cohorts with time and the mean learning growth (the annual increase in mean score) per annum. Treated briefly in a later section are the more complex issues in the relationship of an individual's learning growth with time, the real issue of concern to teachers.

The National Assessment Program-Literacy and Numeracy (NAPLAN) and the relationship of mean learning status of cohorts with time

In Chapter 6, South Australian test data from 1996 through to 2002 are considered. This period includes Years 3 and 5 data and some Year 7 data. Two time points (Years 3 and 5) are insufficient to speculate about the general trajectory of learning from Year 1 to Year 8. Thus the need to establish whether other data sources can provide a basis for estimating a relationship over time, so that a broad model of learning growth can be developed.

The first Australian National Assessment Program-Literacy and Numeracy (NAPLAN) tests were conducted in May 2008 for all students in Years 3, 5, 7 and 9 in government and non-government schools (National Assessment Program Literacy and Numeracy, 2008). This publication has been timely and helpful in the refinement of the understanding of the general trend in test performance with increasing Year levels and age for this thesis. US test norming programs, referenced later, broadly corroborate the general trajectories of mean learning status in reading and numeracy over an extended period of schooling. While the NAPLAN data are cross-sectional, based on Hilton and Patrick (1970), the trends are considered as approximately similar to the longitudinal situation and broadly indicative of the likely trends that existed in the South Australian data of 1997 and 1998.

A detailed consideration of the NAPLAN data is presented in Appendix 5. Appendix Figure A5.1 shows the impact of plotting data points at average age[22] rather than at Year level. Age presentation establishes that age distributed data can be modelled both by a first-degree polynomial and equally well by the Gompertz relation, as advocated by Courtis in Chapter 2. Fitting learning status data by age to a Gompertz curve has some additional benefits over a polynomial fit. These benefits relate to the possible explanations for differential rates of learning, generally and for fast, average and slower developers and are discussed in detail in the appendix. As a result the Gompertz model is used to model the general trajectory of learning with age and Year level when using Rasch model derived vertical scales. The Gompertz model is used in Chapter 6 as the basis of interpolating missing points.

A general model for the NAPLAN 2008 data based on national means

Based on explorations of the fit of the Gompertz relation to the individual state and territory data, omitting the less well performing Northern Territory, a model based on mean age is fitted. Data in Appendix 5 confirm that a model fitted to the national means at average age is virtually identical to one fitted to the more complicated individual State average age points. The model based on national means provides a general indicator of the trajectory for learning status means with age in reading. A similar model can be fitted to the numeracy data. The model uses data from almost all students in Australia in Years 3, 5, 7 and 9 in 2008. The incremental learning with age established in the general model adds support to using the same process to extrapolate from SA data for 1997 and 1998. Most importantly it illustrates that rates of learning growth diminish with age, and that as a consequence snapshots of learning status with age do not sit on a straight line. The Gompertz function, with appropriate parameters, provides a smooth curvilinear trajectory through the data points.

The model development is documented in Appendix 5. *CurveExpert* (Hyams, 2001) software is used to fit a model developed from the Gompertz expression (Gompertz, 1825). The curve fitting is a pragmatic process to idealise the trajectories. Alternative curves can serve this

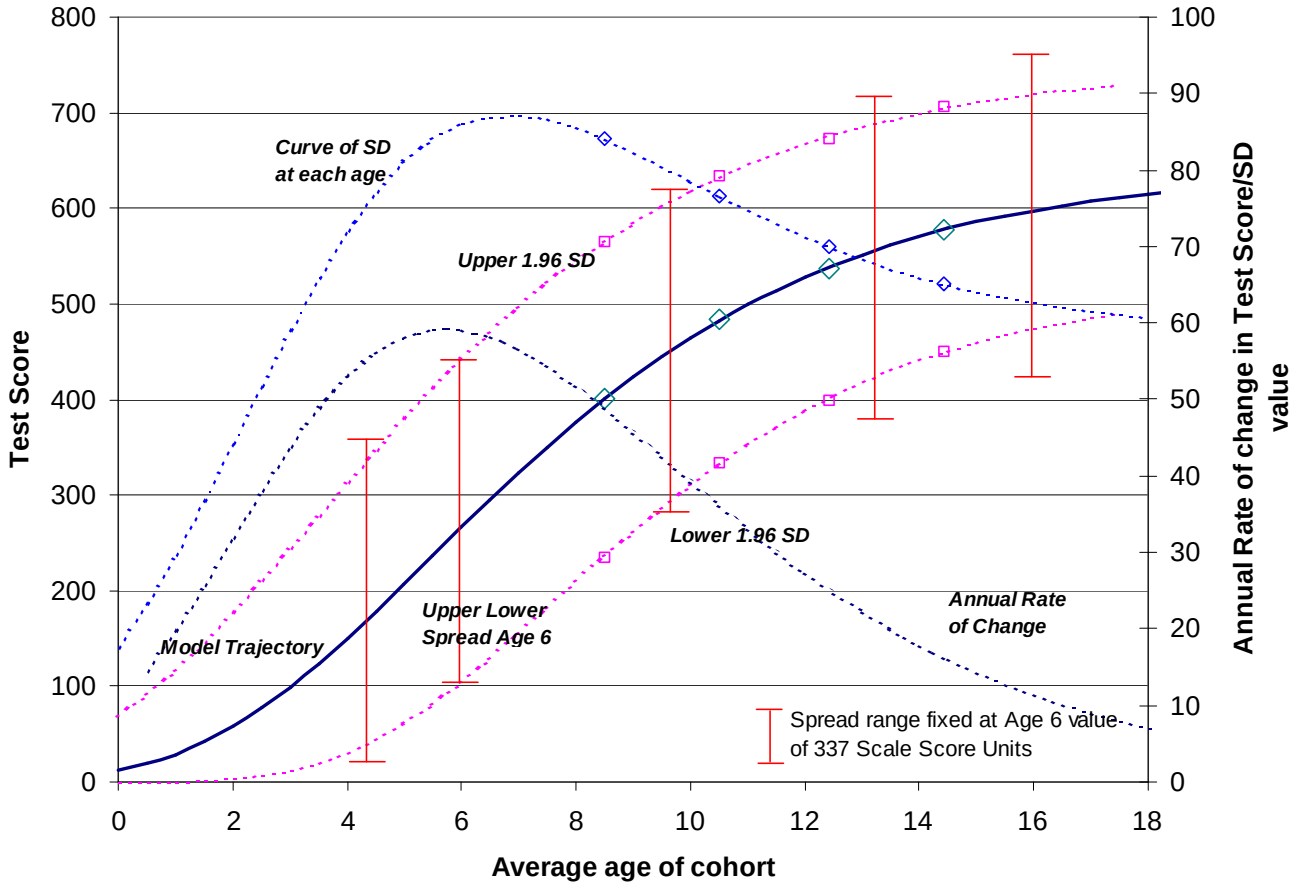
[22] Appendix 5 establishes, as part of the general development of processes to model the general trajectory of learning with time, that comparing Australian system means to an age fitted curve may provide a fairer comparison of performance than the Year level means. System performance is masked when comparisons to the national average are made. Table A5.1 shows that some systems, those with average ages significantly lower or higher than the national average, are misrepresented when compared to the national average. Queensland and Western Australia in particular, while they are both below the national means, sit on the Gompertz model line of best fit for age-spread data. Tasmania on the other hand, having a higher average age than the national average, is shown to be performing less well when age is considered.

purpose. A quadratic curve can provide an approximately equivalent solution (as can some other models applied in biological research included in the *CurveExpert* software). The Gompertz expression is selected for reasons expanded upon and illustrated in Appendices 5, 6 and 7.

The data set used to establish the trajectory of learning with age is very simple; four points only, centred on the national mean ages at Year levels 3, 5, 7 and 9. Fitting a Gompertz model uses an age scale starting at zero. The implication of the value of the mean learning status at age zero is considered in the Appendix. The NAPLAN scale, transformed from a logit scale by the NAPLAN analysts, is set adequately high for the ages 8 (Year 3) to 14 (Year 9) that the scale can take a plausible value at age zero, that is, a NAPLAN scale value of approximately zero. As a result the NAPLAN score value at age zero has little impact on the shape of the curve through the tested Year levels and ages, although it does influence the model trajectory from age 0 to 5. The model development considers the effect of removing any one of the four points. As long as the highest and lowest points are retained, the fitted lines are virtually identical whether or not the intermediate points are included (see Appendix 5, Figure A5.2).

Figure 5.2 illustrates why the Gompertz model is attractive as a model for the general trajectory. The asymmetrically positioned inflection point offers a possible mechanism for what may be happening to reading development with age. Accepting the sigmoid shape and the asymmetric Gompertz curve as appropriate choices to model the trajectory, a mechanism for the rate of learning is provided. The curve of the annual rate of change at each point on the trajectory is shown and scaled on the right axis. The rate is increasing as the inflection point (at about age 6) is approached from the left. On the curves fitted to the 2008 reading data, the rate of learning is increasing rapidly from ages 3 to 6, peaks at about age 6 (at about 66 scale points per annum) and then reduces from ages 6 to 15.

Figure 5.2 Model of NAPLAN Reading 2008 with indication of spread of data



The model also estimates the annual rate of change in reading development at each point on the age scale. The points at 1.96 SDs above and below the means encompass 95% of the student scores at each age. Curves can be fitted to the four actual points that delineate these upper and lower boundaries of the 95% of cases derived from the published SDs (assuming a normal distribution of the learning status means at each age point). The upper curve has an asymptote at 751 on the test scale and a notional test scale intercept at age = 0 of 68 test scale units[23]. The lower curve has an asymptote at 521 on the test scale and a notional test score intercept at age = 0 of 0.08 test scale units. By subtracting the model lower boundary values from the upper boundary values at any age point, an estimate of the SD at that age point can be made by dividing the resulting value by 3.92 (2 x 1.96). On this basis, estimates of the SD can be made for any age points, enabling the estimation of the effect sizes for annual growth at those age points.

[23] This intercept could be forced to be 0 and a slightly modified curve fitted. The impact of allowing the intercept to be 68 increases the initial spread and thus the initial SD estimate.

The resulting SD estimates are plotted with their scale on the right hand axis in Figure 5.2. The estimated SDs start small, grow to about 87 test scale units at about age 7 and reduce from that point on. Based on the observation of Schulz and Nicewander (1997) that growth spurts (e.g., puberty for human height) lead to greater variance at the spurt point, it is confirmed in this model that the SDs are greatest around points of rapid growth, that is near the inflection point.

Taking age 6 as an example point on the age axis, the model estimated learning status is about 270 score points. At this age the maximum annual rate of change for the average student is shown to be about 60 score points per year (right axis). The SD of the spread of scores at age 6 is approximately 85 score points (right axis), just below the peak SD at about age 7. The model estimates a learning status value and a SD value for each age point. A six year old at the 97.5 percentile point has a score around 450 and one at the 2.5 percentile point a score of about 100.

The peak SD lags the peak rate of learning development by about a year. The peak rate of learning for the average student is around 6 years, the peak SD around 7. In this model, logical and mathematical reasons are provided for the scale shrinkage (Yen, 1986; Camilli et al., 1993), the shrinkage of SD within a year level (very small) and the more obvious reduction of SD at higher Year levels relative to lower levels.

The estimated annual rate of learning at each age is also plotted on the right hand scale. This curve illustrates an implication of the model. Students who sit near the trajectory of the mean, that is average students, are likely to be learning at their maximum rate about age 6. The implication of the model for early childhood learning is that the peak rates of learning vary considerably. Those students who are further away from the mean have different ages of peak rate of learning. These are illustrated later in Figure 5.3.

The actual data points for the upper and lower bounds are identified on the curves in Figure 5.2. Also plotted are the actual scores on the model trajectory and the actual SDs. All fit well on their respective curves.

The bars shown on the chart are all of constant length. These are based on the estimated SD at age 6, and indicate visually the reducing spread of the scores at higher ages. Patterns below age 6 are very speculative as the test process cannot be applied below age 7. However the diminishing SD (the narrowing of the spread around the line of the trajectory of the average student) is plausible, as the rate of growth is smaller and the actual quantum of learning that is possible is less. Applying a better basis for estimating the range of pre reading skills would allow the development of a better model. The author has allowed the trajectory to start at age=0 for completeness. At some future point, based on a better recording of the learning of

the appropriate skills in younger children on the test scale, the actual curves could be established and the utility of the model below age 5 tested.

The resulting general form of a model based on the Gompertz expression can be fitted to the actual data points very well, and can be extended through fitting curves to the upper and lower 95% limits of the spread of the scores. In the next chapter the general strategy of fitting Gompertz equations is used with NAPLAN data and the SA data collected in 1997 and 1998 to develop the trajectory for the mean score at each age. The treatment here illustrates that the model offers more value than just the imputation of missing data. It has the potential to explain some aspects of the rates of learning with age as part of the knowledge base for teachers.

The implications of the model can be explored further by plotting the annual rates of learning for the mean, the upper and lower boundary curves and comparing the rates of learning for each curve. (An example can be found in Appendix 5, Figure A5.6). While the model illustrates in an approximate way what data might look like if, say, all students were assessed at the one point in time and their data plotted by their age at testing, the model can also estimate what the mean score for a cohort at a particular cohort average age might look like. Using the model in this way enables the effect size for usual year-to-year growth to be estimated. The next section makes such estimates and compares them to US data to check whether the behaviour of learning in reading in Australian schools is approximately consistent with patterns elsewhere.

Effect sizes for annual growth

The model in Figure 5.2 can be used to estimate likely effect sizes in learning growth from one year to the next. This is helpful in establishing whether the phenomena of decelerating rates of mean leaning status by age/Year level are peculiar to Australia or are general when learning is measured on a vertical scale.

Hill, Bloom, Rebeck Black, and Lipsey (2007) analysed norming data of 7 national US reading tests and 6 national mathematics tests to establish the trend in annual learning growth on vertical scales for each of these tests. Their purpose was to provide “expectations for growth or change in the absence of an intervention” (Hill et al., p. 2) as general benchmark indicators of the effect size required for an intervention at any Year level to be deemed to be greater than expected normal growth. Annual growth in achievement was estimated by taking the difference of mean scale scores in adjacent grades. The difference was converted to a standardized effect size by dividing it by the pooled SD for the normed data in the two adjacent grades. The mean effect size over all tests was then calculated. The results are shown in Table 5.1 in columns (3) and (5).

A similar process is applied to the NAPLAN model in Figure 5.2. The estimated SDs are used to calculate Year level to Year level growth effect sizes. Population sizes are estimated from known values for the four tested cohorts (n ranges from 262,000 to 265,000).

The resultant estimates of effect sizes for NAPLAN reading are also listed in Table 5.1. Two estimates are provided in columns (1) and (2). Column (1) is based on the expected average age at testing. Column (2) is the estimate 6 months later. This second estimate is provided to illustrate the general reduction in effect size as age increases. A shift upwards of 6 months in age reduces the effect sizes by 0.01 to 0.04 SDs per annual growth effect. Effect sizes follow the same trend as the general increments in growth, diminishing as Year level increases.

Table 5.1 Estimated effect sizes for annual reading growth based on the model for NAPLAN trajectory –compared with US effect size estimates for Reading and Mathematics.

	(1) Estimated from NAPLAN Reading model (at mean age at test)	(2) Estimated from NAPLAN Reading model (at mean age at test plus 6 months)	(3) Estimated from 7 pooled US Reading tests	(4) 95% CI US data	(5) Estimated from 6 pooled US Mathematics tests	(6) 95% CI US data
K to 1	0.75	0.74	1.52	(+/- 0.21)	1.14	(+/- 0.22)
1 to 2	0.71	0.68	0.97	(+/- 0.10)	1.03	(+/- 0.11)
2 to 3	0.65	0.61	0.60	(+/- 0.10)	0.89	(+/- 0.12)
3 to 4	0.58	0.54	0.36	(+/- 0.12)	0.52	(+/- 0.11)
4 to 5	0.50	0.46	0.40	(+/- 0.06)	0.56	(+/- 0.08)
5 to 6	0.42	0.39	0.32	(+/- 0.11)	0.41	(+/- 0.06)
6 to 7	0.35	0.31	0.23	(+/- 0.11)	0.30	(+/- 0.05)
7 to 8	0.28	0.25	0.26	(+/- 0.03)	0.32	(+/- 0.03)
8 to 9	0.22	0.20	0.24	(+/- 0.10)	0.22	(+/- 0.08)
9 to 10	0.18	0.15	0.19	(+/- 0.08)	0.25	(+/- 0.05)
Mean Effect Size	0.47	0.43	0.51		0.56	

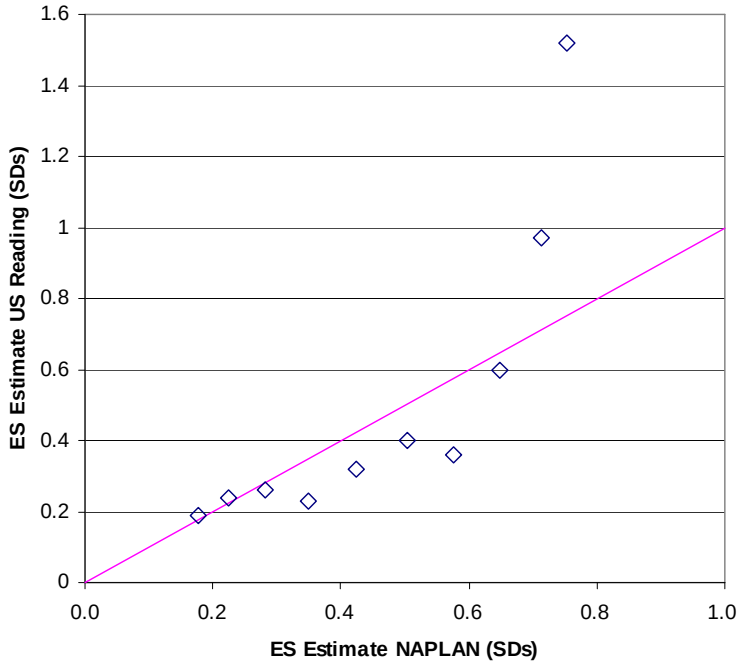
US Data: Table 1, Hill et al., 2007.

Test Sources: Annual gain for reading is calculated from seven nationally normed tests: CAT5, SAT9, TerraNova-CTBS, MAT8, TerraNova-CAT, SAT10, and Gates-MacGinitie. Annual gain for math is calculated from six nationally normed tests: CAT5, SAT9, TerraNova-CTBS, MAT8, Terra Nova-CAT, and SAT10.

General trends for the NAPLAN model and for US estimates are similar. Reading effect sizes for the two sources are plotted on perpendicular axes in Figure 5.3. The points are near to and spread along the identity line. Rates of growth are markedly higher for lower Year levels in the US data relative to the NAPLAN estimates but the same general trend is confirmed. However as the effect sizes for K to 1 and 1 to 2 are much higher in the US, this implies a steeper rate of growth (as measured by the US tests) than in the NAPLAN model. This might imply that the NAPLAN model developed above, underestimates the rate of growth in the

lower ages, that is the current Gompertz model parameters are conservative in the estimate of growth rate (in the unmeasured students below age 8). As illustrated in Appendix 5, the steepness of the trajectory (and the rate of learning) can be influenced by adjusting the assumed position at age=0. The NAPLAN model described may not reflect the real rate of early learning.

Figure 5.3 Comparison of effect sizes at each Year level-NAPLAN, US



Mean effect sizes, averaged over all Year levels are also reported in Table 5.1. Mean values are in the range 0.43 to 0.56, which compare well with the estimate by Hattie that the "average or typical effect of schooling was 0.40 (SE = 0.05), providing a benchmark figure or 'standard' from which to judge the various influences on achievement" (Hattie &Timperley, 2007, p. 83). Hattie (1999) estimates the effect of a year of schooling as being 1.0 SD (Hattie, 1999, p. 4), greater than the estimates in Table 5.1.

The refinement that is possible in the estimate of base effect size from the above NAPLAN model analysis and the Hill et al. (2007) analysis relative to the Hattie estimate, is the pattern of variation in this average effect size with Year level. At lower levels much greater intervention effects appear to be required to show an effect greater than the general underlying trend in rate of learning at these levels/ages. Averaged over all Year levels the general estimate of Hattie (0.4) is comparable although his speculated value for annual growth would appear to be overestimated.

The NAPLAN data are the full national population trends. Understanding the general trends in learning with Year level and age is one of the possible benefits of this comprehensive

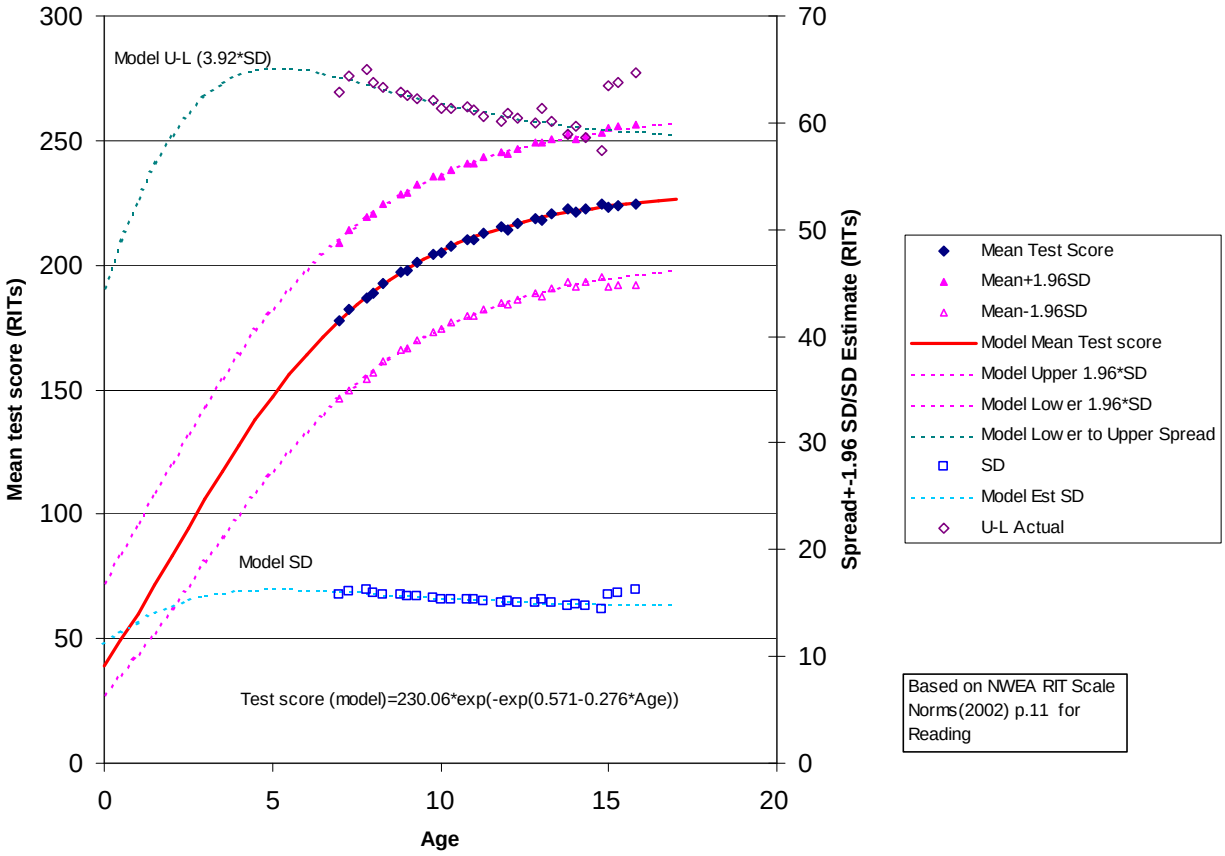
testing program. It is remarkable how a simple and useful model can be estimated from 4 main data points and their SDs. Access to student level records with the ability to monitor the longitudinal growth of individual students from Year 3 to Year 9 should lead to even more refined and interesting models. A model derived from the NAPLAN data would provide one more element in a teacher knowledge base to help classroom teachers place their own data in context.

A complementary example from a US source follows, with a similar general model being developed. This is done to confirm the general trajectories, already partly confirmed through the Hill et al. effect size analysis.

Annual growth in the Northwest Evaluation Association (NWEA) norms and effect size estimates

Sources of longitudinal and cross-sectional data of test scores by grade are not readily found in the public domain. A rare source of such data is the Northwest Evaluation Association (NWEA), a not-for-profit organization operating since 1977, which provides assessment products and services to US schools, school districts and states. Data amassed over more than 20 years provide measures of student learning growth. More than 3 million students have been assessed through NWEA, which has established a rich database of student assessments. NWEA use a logit-based measurement scale that has been confirmed by regular evaluation to be stable and valid over time (Kingsbury. 2003; McCall, 2006). The vertical scale is developed using the Rasch model. As described earlier in Chapter 1, the Rasch model allows alignment of student achievement levels with item difficulties on the same scale. The scale is calibrated in RITs (abbreviation of Rasch Unit coined by NWEA) and is a transformation of a logit scale, such that 10 RITs = 1 logit.

Figure 5.4 NWEA Reading Norms data (2002) with fitted curves



Source: Northwest Evaluation Association (2002)

Figure 5.4 is developed in Appendix 6 and is based on the NWEA 2002 norms for reading. The data sets are cross-sectional but with points within a grade longitudinal. The data points come from three assessments per grade (one interpolated), each positioned at an estimated average age at testing as described in the Appendix.

The plot corroborates the general curve of the NAPLAN data, that is decreasing growth in learning with age but with many more data points to add certainty to the general shape.

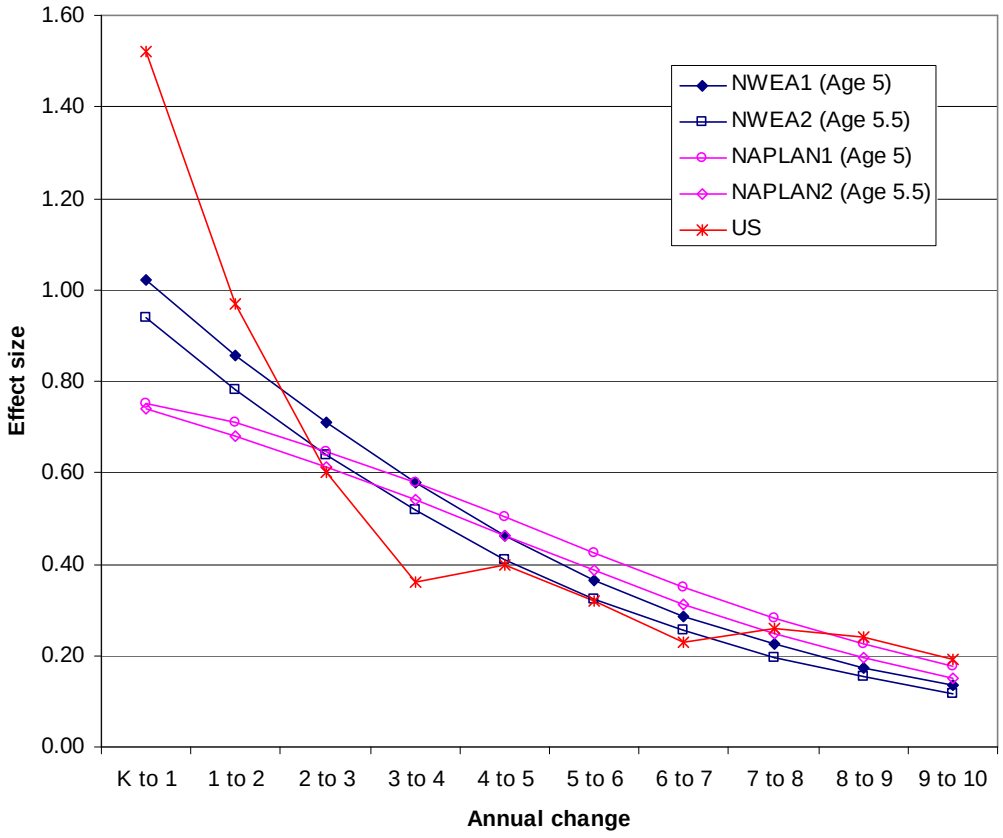
Points for fitting the upper and lower boundary lines are calculated from 1.96 x SD and the curves fitted independently. A model for the 2.5th to 97.5th percentile spread can be developed from the modelled upper and lower boundaries by subtracting the difference for each age point. The resultant Upper minus Lower curve tracks the actual spread quite well, except for the points at age 15 (which were based on a much smaller sample and thus estimated with higher measurement error). The model Upper-Lower spread is greatest near the inflection point, around age 5, where annual gain (rate of learning per annum) is greatest.

The SD can be estimated from the spread model by dividing by 3.92 (2x1.96). The result is a model for the SD that appears plausible. SD is at its greatest at about age 4 to 5, and tracks

through the actual SD data points quite well (except for age 15 as mentioned above). This fits once again with the Schulz and Nicewander (1997) observation of growth spurts and increased spread referenced previously. Apart from at Age 15, it is consistent with the scale contraction effect (Yen, 1986; Camilli et al., 1993) discussed in the NAPLAN model.

The modelled trajectory for the test scores by age and the modelled SDs enable effect sizes for annual growth to be estimated, using the sample sizes in the original NWEA norming data. The resulting effect sizes for ages 5 through 15 (K through 10) are illustrated alongside the NAPLAN and earlier US estimates in Figure 5.5.

Figure 5.5 Effect size estimates for NWEA, NAPLAN and general US norms for Reading



The effect sizes are estimated at two assumed ages for the Kindergarten year. From '3 to 4' onwards, the location where actual data points exist for NWEA and NAPLAN, trends in effect size are similar (even though actual effect sizes vary). The modelled effect size trends from K to 3 are more divergent as a result of the differences in the modelled trajectories. However the composite US trend, based on data from a range of tests, indicates much higher effect sizes than either of the suggested models. The overall conclusion is that growth patterns that influence effect sizes are very similar, with age or grade. The annual growth values obtained by comparing successive grade means of learning status diminish systematically and the trajectory of the learning path is not linear.

Mathematics Assessment for Learning and Teaching (MaLT) in England

The same general trajectories also apply for mathematics. Williams, Wo, and Lewis (2007) in the development of a mathematics assessment in the UK provide a final confirming example of the non-linear growth of learning. Williams et al. (2007) and Ryan and Williams (2007) report data from a national sample designed to provide age related performance references for the MaLT test. Year level cohorts of between 1000 and 1400 students were recruited from 111 schools.

Data are summarised by the developers with the time dimension calibrated in months. The test was developed using the Rasch model. Vertical equating was through common persons across Year levels (about 1/3 of the cohorts sat adjacent level tests). Common item equating was applied in the test development phase where about half the items for the next Year level for pre-test cohorts were included in the lower level (Williams et al., 2007, p. 132). The derivation of the model is described in Appendix 7.

Figure 5.6 Model of Mathematics Development - Mathematics Assessment for Learning and Teaching, (MaLT)

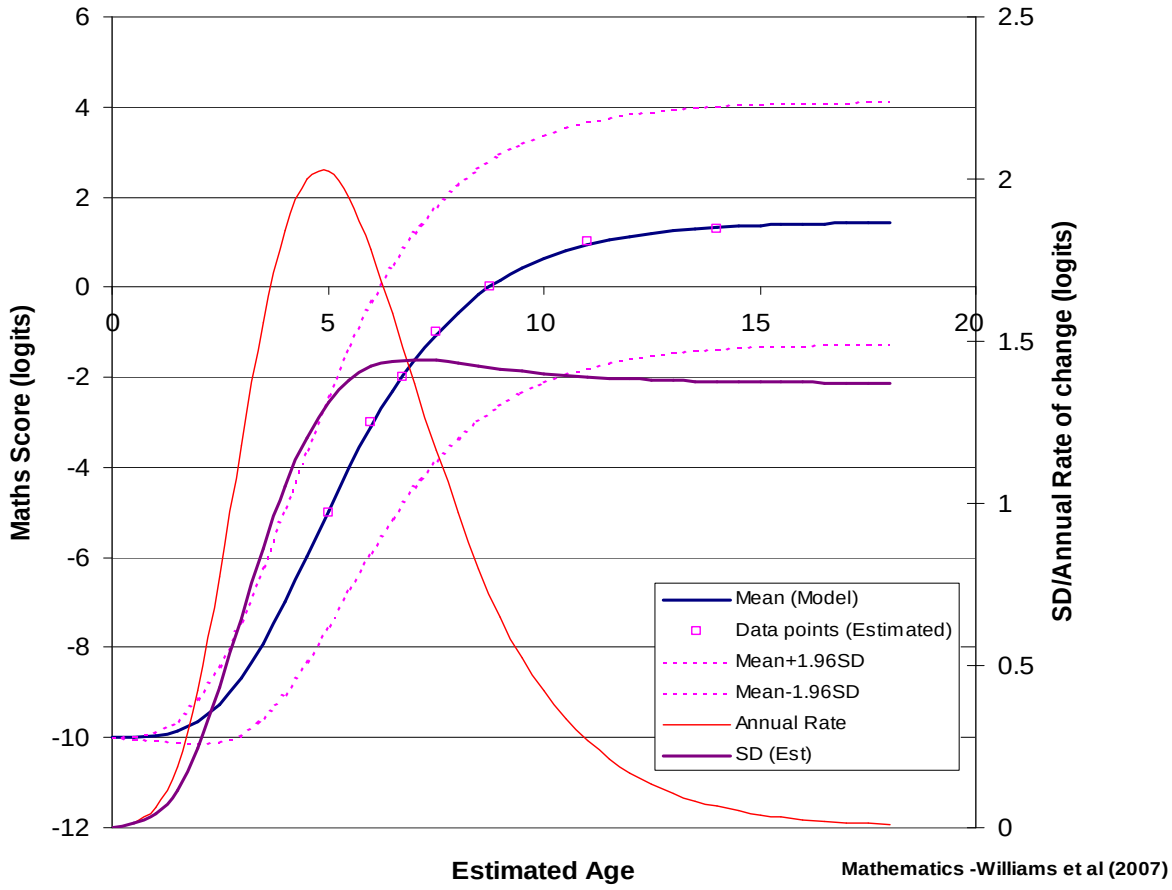
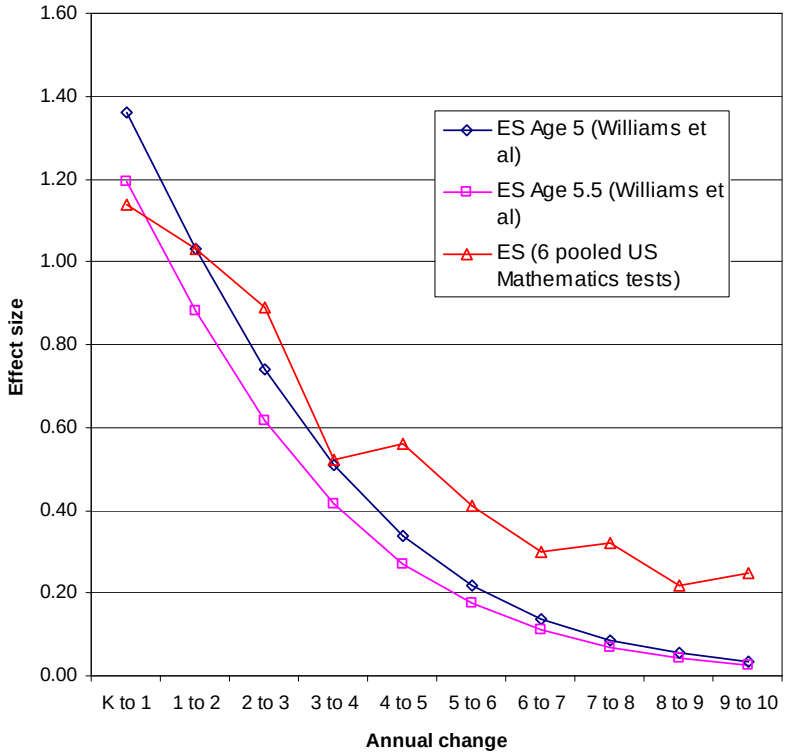


Figure 5.6 displays the resulting model for the mean, the actual data points and the estimated upper and lower boundaries for 90% of the data. Also plotted is the annual rate of change based on the model with its scale on the right hand axis. The estimate of the SD is also

plotted with its scale on the right hand axis. Consistent with the NAPLAN model, the model SD reduces slightly as age increases and peaks about a year of age past the inflection point. As previously, model estimates can be used to estimate effect sizes for year-to-year growth. These are shown in Figure 5.7. For reference the US effect sizes from Hill et al. (2007) for 6 Mathematics tests are included (listed in Table 5.2 above).

Figure 5.7 Effect sizes for Mathematics Assessment for Learning and Teaching compared with pooled US tests



The general pattern of reducing growth in learning mathematics is exhibited, through reducing effect sizes, in both the Williams et al. and US data. Somewhat surprisingly, for the model for England (Williams et al.), the growth in logits per annum from Figure 5.6 and the effect size in SDs from Figure 5.7 are both close to zero by the transition from Year 8 to 9, much lower than in the US comparison. (Year 10 Williams et al. data are extrapolated using the model; no data were collected at Year 10). It was this plateau effect that was the focus of Williams et al. (2007) since it implies almost no mathematics development from Years 7 to 9.

The validity of the vertical scale is considered in Williams et al. (2007). With the qualification that the phenomenon might be related to the inadequacy of the scaling, Williams et al. conclude that

> It seems realistic to conclude that progress is indeed very slow (about 0.2 logits per year) over this period. … One speculates that the repeated exposure to the same curriculum in secondary school has a negative effect on these common learning outcomes. (Williams et al., 2007, p. 139)

The Williams et al. data are also helpful in illustrating the age effect within a Year level cohort. This phenomenon appears to apply quite generally and is covered in a later segment of the chapter.

General conclusions on learning growth trajectories from cross-sectional data

Using the data from the three cases cited, smooth curves can be fitted to describe the trajectories of the means of Year level/age groups, with Gompertz models providing adequate fit in each case. Quadratic models also fit well but do not provide the same potential for hypothesis development. Learning development over time is non-linear in all of the above cases.

Data from other US sources (Hauser, 2003 for NWEA data; Northwest Evaluation Association, 2005, for RIT norms; Williamson, 2006 with Lexiles; Walston, Rathbun, & Germino Hausken, 2008 and Pollack, Atkins-Burnett, Najarian & Rock, 2005 with the Early Childhood Longitudinal Study; Star Reading, 2005 for Star Reading norms) confirm a generally common pattern of growth for vertically scaled tests. The rate of growth is greatest in the early years with year-to-year growth diminishing with Year level (or its direct equivalent age). Effect size estimates of Hill et al. (2007) confirm the general diminishing learning growth with age and Year level.

There are exceptions to the non-linear growth with Year level. Reports by Rothman (1998, 1999), Rowe and Hill (1996) and the Victoria CSF/VELS (and Chapter 7 in this thesis) using teacher judgement data show straight-line growth with Year level. A linear relationship with grade is often indicative of a grade equivalent rescaling (Schulz & Nicewander, 1997). This insight may offer an explanation for what teachers are doing in their level scaled teacher judgement assessments. Perhaps they are basing their assessments on an internalised grade equivalent standard that can be expressed using the levels scale. This possibility is addressed later.

In summary, smooth curves can be fitted for most learning areas where the Rasch model has been used to develop the learning scale. Learning growth in these vertically scaled examples is non-linear. All cases draw on for longitudinal data show a diminishing growth rate for learning in specific learning areas with age/higher Year levels where Rasch scaling is applied. Whether this apparently universal phenomenon is ‘normal’ or due to poor curriculum structure, poor pedagogy or other factors is an open question. There is no doubt that cognitive development and thinking skills can be ‘accelerated’ (Adey & Shayer, 1994; Endler & Bond, 2007). Figure 2 (Endler & Bond) in particular, while showing the universal curvilinear form for the control group also show that cognitive acceleration occurs in particular pedagogical treatments. The general shape of the trajectory of learning however,

while elevated relative to the control group, still appears to show a diminishing growth rate along a logit scaled axis.

For reading and numeracy the Gompertz model provides an adequate mathematical description for the trajectories of cross-sectional cohorts. The Gompertz model tends to an upper asymptote (which seems logical since the scale is based on difficulty) and describes a trajectory in the ages below 8 in a form that has an attractive heuristic logic, including implying a peak rate of learning at about age 6, based on scale assumptions. Other complementary evidence later in this chapter will illustrate the steepness of the initial learning from age 5 to 7. The Australian data, when expressed as effect sizes, are comparable to the mean effect sizes of the grand mean of a large number of the vertically aligned US tests in reading, confirming that the pattern of growth, grade to grade, exhibited by cohorts of US students is also non-linear. The evidence suggests, in broad terms, the same general trends apply in Australia, England and the US.

The general Gompertz model provides a pragmatic process for modelling group means as well as suggesting some areas for further hypothesis development (rates of learning, SD trends among others) that can be tested. Generally, learning growth can be modelled with asymmetric sigmoid functions leading to a decelerating rate of growth past the inflection point.

Further understanding of what is happening in learning growth can be obtained by exploring finer resolution age groupings within a Year level. This understanding provides a basis for the extrapolation of data within a Year level. It also provides an unanticipated benefit, a characteristic shape of learning development within a Year level cohort, which can be used as an indicator of test-like data when teacher judgement data are being examined in Chapter 7.

Patterns by age within Year level

Test score means for a Year level (or grade) cohort, when spread by the relative ages of the students, have strong identifying characteristics, sometimes described as the ‘birthday’ effect. Test scores for a Year level cohort of students generate a characteristic shape for the average scores of students by age, where the age categories are made finer, for example in months from birth or decimal age (8.2, 8.3 etc.) at the date of testing.

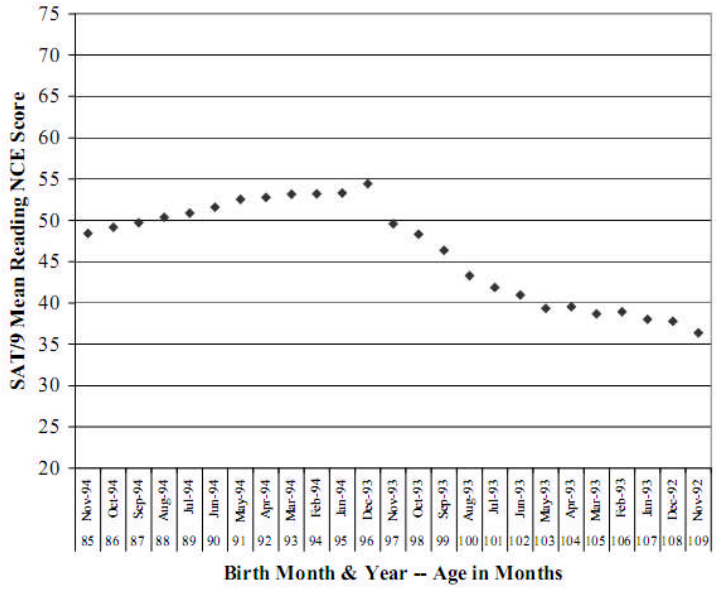
Cahan and Davis (1987) show an increasing percent correct score for reading and mathematics within a cohort with age (in months) for Israeli children. The same test was applied in Year levels 1 and 2, and the increasing score within grade by age was consistent in both grades. The scores for students either older or younger than the appropriate ages for the grades were not reported.

Grissom (2004) reports a wider set of ages within grade and for a large number of students for California. For reading the mean scores by age in months, in Grades 2, 6 and 10 are reported. Cohorts range from 388,000 (Grade 10) to 455,000 (Grade 2). For mathematics the mean scores by age in months for Grades 2 and 6, with similar cohort sizes are reported. The tests in all cases were versions of the SAT 9.

The pattern of the means by age is very similar in all cases. Figure 5.8 (from Grissom, 2004, Figure 1, p. 6) is typical of the general shape, in this example for reading. The left section of the figure shows the increasing mean scores for the students in the normal age range for the grade. Once the highest age for the normal age range is reached, the mean scores decline as shown in the right section of the figure. The growth in mean score from the youngest group to the highest within normal age group is 6 score units for reading. The pattern is consistent for Grades 6 and 10 with the youngest to oldest difference reducing to 4.5 and 1.7 score units respectively. The age effect continues to Grade 10 but is markedly diminished.

The same effect applies for the mathematics scores in Grades 2 and 6. Youngest to oldest difference within the normal 12 month age range for the grade in mathematics at Grade 2 is estimated to be about 8 score points, reducing to about 5 for Grade 6. While the effect is clear the spread of scores at each age is very wide. Accordingly the age difference explains only a very small component of the variance. The value to this thesis is the consistency of the pattern across grades. This pattern is a potential marker of what a test applied within a grade typically generates as a pattern by age.

Figure 5.8 SAT 9 Reading Scores Grade 2 (2002)- from Grissom (2004, p. 6)

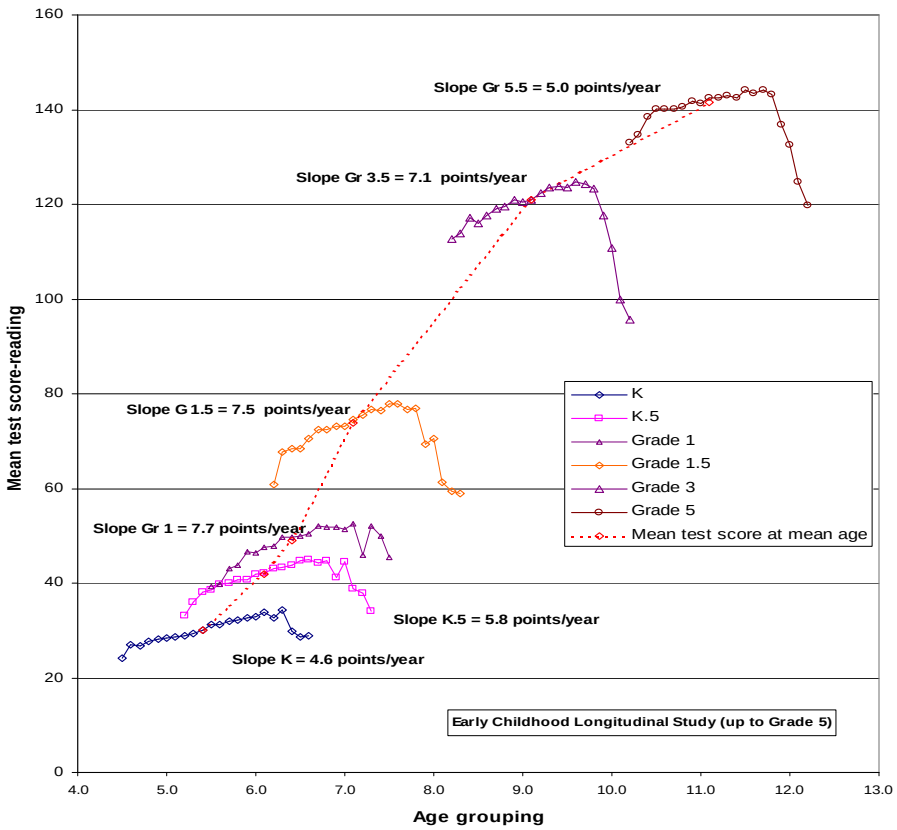


Williams et al. (2007) report the same general learning growth with age within a Year level cohort in the norms developed for the MaLt project described earlier. The gradients within

the normal age for Year level segments generally show higher mean scores for each progressively older month grouping. The sample size per month is small, approximately 100 students. This leads to some variability on the phenomenon but the general trend is an increase in the mean score for the group for each month of age.

The US Early Childhood Longitudinal Study (ECLS) public data set for a national sample of over 11,000 students over six assessment periods (Pollack et al., 2005) is summarised below. The data are summarised from the original data sets (Early Childhood Longitudinal Study, 2004; Tourangeau et al., 2006). In this study students were assessed from the 1998-1999 school year in Kindergarten through to spring 2004 for Grade 5 (with a final assessment cycle in 2007 in Grade 8 only partly published at the point of writing). The mean test score in reading, on a vertically scaled assessment scheme is shown in Figure 5.9, for groups of students by estimated age at testing.

Figure 5.9 Reading Test scores Early Childhood Longitudinal Study (ECLS) by age at testing



Each point represents a group with a common decimal age. The mean score increases for each group until the range of the normal age for grade is reached. At this point there is a sudden drop in the number of students and a marked drop in the mean score as age increases (with diminishing numbers of cases in this tail). Each of the normal age groups has a sample size of about 900, dropping off after the peak mean score to less than 100 students. The shapes of the curves for each panel are similar as age increases, but with the tail becoming

longer as Year level increases. Each gradient of the slope of improvement with age is marked on the graph. The gradient starts at 4.6 points per year of age in Kindergarten (effectively 0.46 points for each 0.1 of age), peaks in Grade 1 at 7.7 points a year and then reduces gradually to 5.0 points by Grade 5 (reasonably consistent with the model in Figure 5.2).

The dotted line connects the mean score at mean age for each cohort. The trajectory accelerates from Kindergarten to Grade 1, and then reduces gently as Year level increases, consistent with the general trajectories described earlier in the chapter. The prominent feature of the graph is the increasing mean score by decimal age within a Year level and the sudden drop off once the normal age range of the cohort is passed. This pattern is consistent with Grissom (2004).

This shape is proposed as a benchmark comparison for the South Australian test data and more importantly as an indicator of the degree to which teacher judgement data also display this feature.

Effect size for each 0.1 of age is approximately 0.05. Over a period of 1 year of age this becomes 0.5, comparable to effect sizes quoted earlier. Data reported here for ECLS are up to Grade 5 only (although by 2009 this had been extended to Grade 8) with the effect sizes expected to diminish as Grade increases. In the Hill et al. (2007) summary referenced earlier in Table 5.1, the effect size from 4 to 5 for reading was 0.46 and the NAPLAN model was 0.5 in the same age/grade region.

Further examples of the within-Year level phenomenon are found in Bedard and Dhuey (2006) who show the effect occurs in 19 countries based on an analysis of TIMMS data. Crawford, Dearden, and Meghir (2007) analyse Key Stage 1, 2 and 3 data for England for a number of years with up to 1.5 million cases for Key Stages 1, 2 and 3 in longitudinal panels. The same effect is found at all Key Stages, although, consistent with the Grissom analysis, diminishing at higher stages. Strom (2004) confirms that Norwegian 15 and 16 year olds, in PISA 2000 data, show the general pattern of improved reading performance with age within a grade cohort. The pattern by quarters of a year (3 months averaged together) is clear; the month by month summary shows two aberrant points (at 5 and 7 months) but sample sizes in the Norwegian PISA sample for a month are small (about 300 students), and thus have a greater standard error of the mean at this level of disaggregation.

The idealised shape of the curve of the mean test-scale-score at each point of decimal age, takes the form of an elongated incline with a tail (or the reverse depending upon the convention for the age axis). Within the age appropriate zone the average score increases until the last age appropriate category. Then the average score decreases again.

The age profiles of test scores within a Year level for tests become the equivalent of fingerprints or DNA markers that typify what test data might look like for a cohort of students in specific learning areas. If such identical fingerprints are also found in data generated by teacher assessments in a common population of students, another form of comparability of the two assessment processes could be confirmed.

The chapter concludes with brief considerations of two issues. The first is an illustration of possible sources for building a teachers' knowledge base as required by Fullan et al. (2006). The case studies confirm the value of tests in understanding learning progressions and providing methods to create learning records for individual students. The second issue is that of the complexity of individual student learning trajectories. These are more varied and less predictable than the trajectories of groups.

Case studies where further analysis of test data might provide scaled indicators of student development

Two case studies are reported briefly. The first (Appendix 8) takes advantage of data that is collected automatically as part of a large assessment support function provided to subscribing schools by the Curriculum, Evaluation and Management (CEM) Centre at Durham University. The centre provides an individual student assessment at each Year level from Reception through to Year 6 and has built up an extensive database of assessments for about 300,000 primary students per annum (CEM PIPS Newsletter 24, 2008). This database enables longitudinal research as well as other forms of data exploration. The assessment format is also applied to subscribing schools in Australia, New Zealand, China and in a range of International Schools. Appendix 8 uses data provided from the CEM to develop a possible learning pathway for the recognition and naming of numerals, one of the first steps in numeracy development. Appendix 8 illustrates that learning orders are essentially consistent across English speaking cultures and that a general order for naming numerals can be empirically determined.

The second case study (Appendix 9) draws on learning progressions developed at the Center for Urban School Improvement in Chicago over a ten-year period. The Strategic Teaching and Evaluation of Progress (STEP) developmental assessment process for reading was created in conjunction with the Chicago Public Schools. This case study illustrates the utility of empirical evaluation of item difficulty in highlighting the steps/stages children go through in developing their reading skills. The example illustrates in particular the likely orders for learning to recognise and name letters of the alphabet in their upper and lower case forms, as well as the order in which letters can be paired with their sound.

In this thesis when considering the possibility of teachers generating assessment data directly, the numeral and letter orders provide the scale for teachers to observe the very subtle natural development of these skills. Assuming no deliberate coaching in out-of-order letters or numbers, the recognition skill displayed by a student at any time is likely to indicate the learning status.

The scale also indicates the relative difficulties of the easiest to recognise characters to the most difficult. For numerals the span from the easiest single digit number to the hardest three-digit number is about 10 logits. This is illustrated in Figure 5.10. The single digits are learned almost in numerical order, except for 7 being slightly easier than 6 and 9 being harder than 10. Two digit numbers do not follow numerical order. Below 20, 13 is the most difficult to learn, more difficult than 20. The first 20 numbers (1 to 13) span over 5 logits. The three-digit number 100 is around the same difficulty as the mid range two digit numbers. Changes in difficulty are very small once the key first 20 numbers are learned. For letter naming and letter sound recognition (Appendix 9) the span from easiest to hardest is 7.5 logits. This is illustrated in Figure 5.11. O, whether upper or lower case, is the easiest letter to recognise. Lower case q is the most difficult and one logit harder than the next hardest, lower case g.

These scales provide a basis for monitoring the development of these critical early character recognition skills but they also highlight the perhaps unappreciated difficulty for students in achieving these first skills. The scale value for a developed letter or numeral could provide a basis for recording learning status. While the logit lengths may vary relative to other tests, the logit scores still provide a general indication of the relative difficulty of early skills learning compared with later learning in reading comprehension.

Figure 5.10 Numbers in Estimated Order of Difficulty to Say Aloud-all numbers to 20, samples from thereon (Difficulties relative to '1')

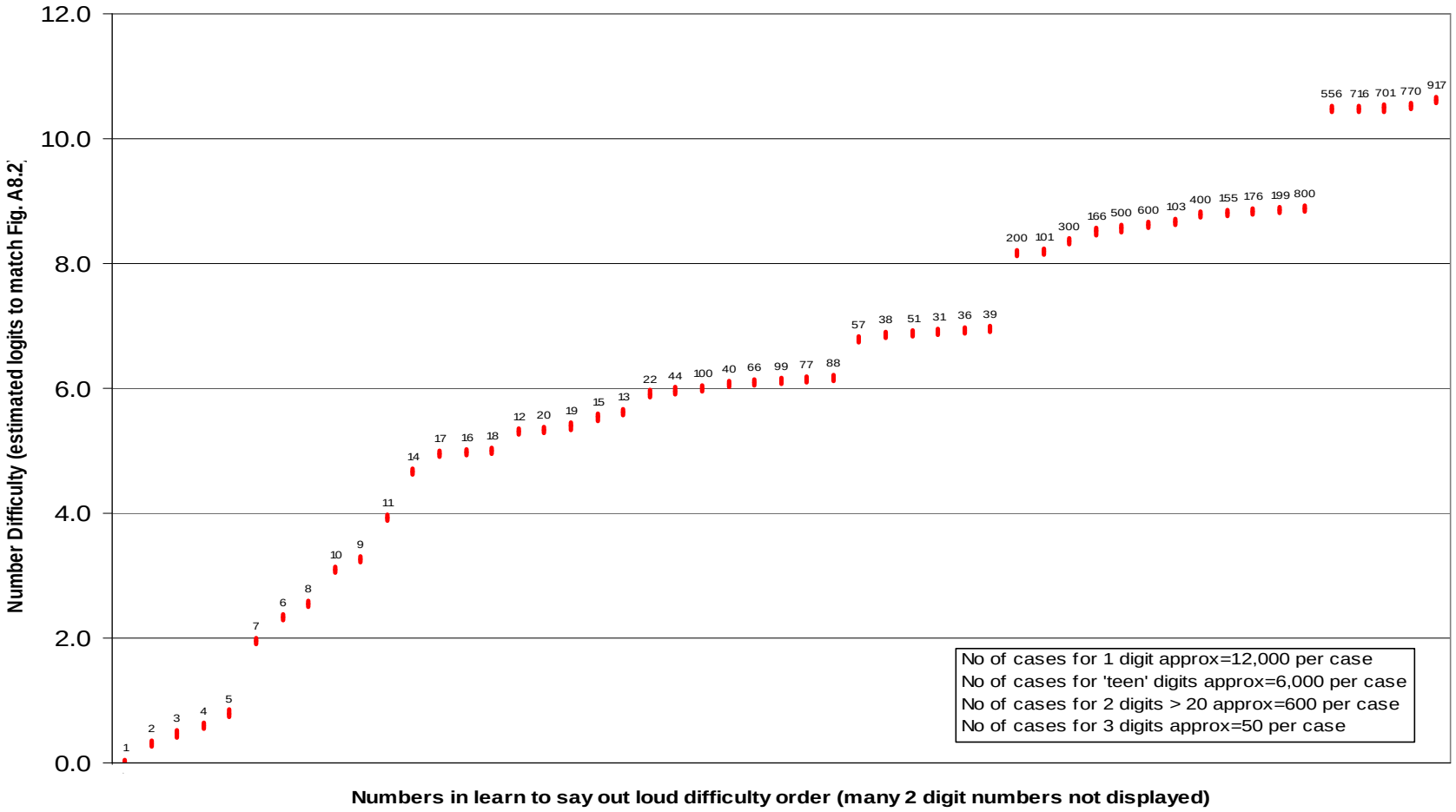
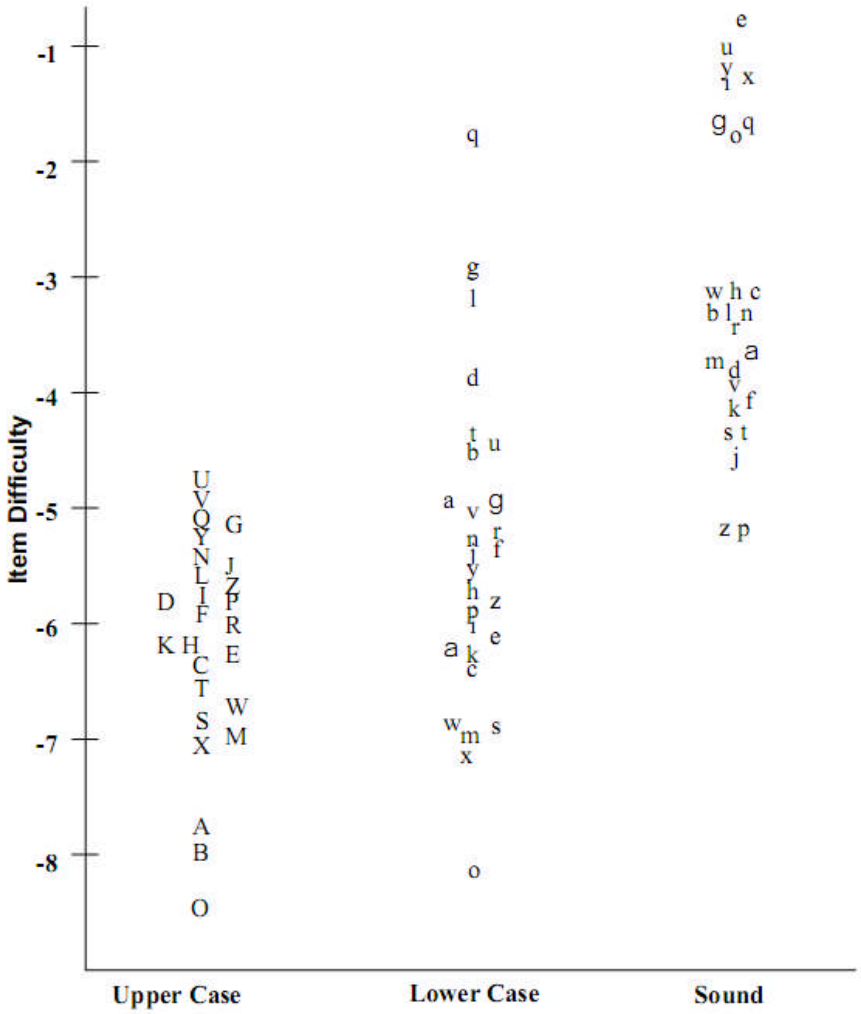


Figure 5.11 Overview of STEP Letter Identification and Letter Sound Item Maps (from Figure 5 Kerbow & Bryk, 2005)



The increase in mean score in logits for the average student from Year 3 to Year 7 is only about 2.5 logits (shown later in Figure 6.1). The steep trajectories of the learning growth for students from ages 4 to 6 in the models developed earlier in the chapter, are consistent with the general difficulty scale reflected in developing character recognition. Any stalling or general (natural) developmental delay in these critical early stages will have a significant impact on the time to develop later skills. That there is likely to be a natural order for learning the names of the letters is confirmed independently by Justice et al. (2006) where their order (based on 339 students only) has a correlation with the order in Figure 5.11 of 0.85.

Understanding the variation in individual learning trajectories, assuming that teacher judgement assessments can be made with finer resolution than appears currently accepted, is important in the context of the thesis. Organised educational assessment practices, although having applied for over a century, are weak in demonstrating fine grain (short time interval)

trajectories for learning. A brief outline of the range of these trajectories is presented next for completeness, as one more element needed for consideration in any system re-design based on utilising teacher judgement assessments.

Comments on individual learning trajectories

The data analysed in this thesis are cross-sectional. They offer only limited appreciation of the potential for longitudinal data for individual students through direct judgement assessment. If reliable longitudinal data were to become available through teacher judgement assessment, an understanding of the range of satisfactory individual student trajectories will be complex for teachers. The general issues are addressed briefly in Appendix 10 rather than as part of the general argument, as they are follow-on concerns after the confirmation or otherwise of the adequacy of teacher judgement assessments.

It appears that time series data for individual students are relatively new data sets. Molenaar (2004) argues the importance of intra-individual variation (IAV) as distinct from variation between individuals (inter-individual variation –IEV), the latter being the major focus of psychology to date in his view. Time-dependent variation within a single participant's time series (Molenaar, 2004, p. 202) is still a very new field of research. A summary of Molenaar's views is provided in Appendix 10.

The essence of Molenaar's argument is that different approaches are required and different results are obtained when one follows individuals, as against aggregates of individuals, over time. This point is made as evidence of the complexity of the problem that teachers would face were more data provided, or developed by them, to follow the learning trajectories of individual students. Based on Molenaar's analysis, any computer support system for the management of learning based on simple extrapolations of individual trajectories from population patterns would be inaccurate. Further recent publications (Molenaar & Campbell, 2009; Molenaar, Sinclair, Rovine, Ram & Corneal, 2009) indicate that there is little literature and analytical support for intra-individual variation modelling:

When students are tracked between two widely separated points in time (K to Year 5), with intermediate values plotted (See Appendix 10, Figures A10.1 and A10.2 as examples), the trajectories can be quite different. Even in the special case where students start with equivalent scores and finish with equivalent scores, the paths taken are varied. Part of the variation in pathway is measurement error. An inaccurate measurement has high impact when only a few data points are possible. More data points, visually presented as graphs, would help identify likely inaccurate measurements.

A major source of the variation is likely to be the idiosyncratic learning process for each student. These idiosyncratic pathways raise large issues for teachers when more data points are available, even if they come from other sources than teacher judgement assessment. They will need to consider when significant changes in learning management strategy for any student are required. The science here is so new, with so little extended longitudinal data, that the initial knowledge base support will be a challenge to develop.

Without adequate analytical tools to make sense of longitudinal data, the benefits to students from more regular records of learning growth whatever their source will not be obtained. New processes for managing these records are required and these must assume a wide range of sources; standardised and online tests, observations, class assessments, embedded assessments. While a graphic history for each student can be displayed to help teachers see each student's development, a technique is also required to identify genuinely stalled trajectories that fall outside the range of normal development. All this presumes that the skill of teachers as 'on balance' judgement assessors can be confirmed and that the monitoring of individual growth trajectories (if made feasible) will help teachers manage the learning support required for each student. It is also assumed that monitoring students in fine detail against a validated scale using a variety of tools will lead to improved outcomes. This is a hypothesis that needs testing.

If the volume of data is to be made manageable, a range of analytical tools to help teachers understand their data will be needed. There are a number of issues that will be relevant in developing these tools. Given that trajectories of learning are idiosyncratic, they may not be able to be projected forward with confidence. The development of analytical models for individual development analysis is in its early days. Group data most likely can be used to estimate only some of the parameters for modelling individual growth. Other parameters will be specific to each individual and derived from their early trajectory. Models based on the previously achieved points and previous estimates of rates of change for the individual are the most useful predictors of the next learning status point at t=x. This is implied in Molenaar et al. (2009) and Malone, Suppes, Macken, Zanotti and Kanerva (1979) and raised in Appendix 10.

Independent of the source of the time series data for each student, the development of the interpretative models to help teachers in the management of learning as students make progress, will be a very interesting challenge. Breakthrough reform anticipated by Fullan et al. (2006) will need many individual times series data sets, with frequent data points on the time axis and low errors of measurement on the learning axes for each strand, to develop the models for the knowledge base.

Summary

The purpose of this chapter was to complete the consideration of themes and ideas seen as elements necessary to build an understanding of the (measured) pathways with time that learners take as they develop reading and mathematics skills. This was addressed for a number of reasons.

The typical trajectories of the means of cohorts as they move through Year levels provide some guides for imputing data to add to the incomplete data set of test results to be addressed in the next chapter. Furthermore models for growth in learning provide some general insights into what might be expected generally as learners move through Year levels.

Understanding the relationship of learning development with age, as described by tests on vertical scales, leads to the recognition of a general age effect in test assessment. This effect is refined when Year levels are analysed separately. A consistent pattern by age within a Year level/grade cohort provides an additional basis for evaluating the effectiveness of teachers in judging the learning status of students, through comparison with age summaries of test data.

The trajectories of the mean can be modelled by a number of fitted curves. The trajectories of individual students are less straightforward than their group means. While only addressed briefly the individual trajectories are shown to vary widely. Techniques to project the forward trajectory for individual students are cutting edge issues in individual psychology. The focus on self-referenced development (intra individual variation) is an open topic with significant implications to education. Such models and projections are necessary to help teachers in their assessments but as well to provide a context for any individual student trajectory.

Understanding learning development and trajectories from large scale testing processes (NAPLAN, CEM, STEP) might be used to provide detail to assessment frameworks and scales for teachers to inform their judgement assessments. With access to potentially rich insights about fine grain learning from that data, monitoring learning directly by observation might be enhanced. Two examples illustrated that useful insights about general learning dynamics can be obtained from test/standardised assessment analysis processes. The two examples show that in the key early stages of language and number learning, what has been learnt (which numbers, which letters) can be indicators of learning progress and relatively easily observed by teachers.

A wide range of matters relating to testing, teacher judgement assessment, the development of levelled curricula, the application of teacher judgment in schools systems and in this chapter, the patterns of growth that assessment data illustrate have been assembled. These matters set the context for conclusions that can be drawn in the final chapter of the thesis. The next three chapters analyse and summarise the specific learning trajectories that can be developed from SA test and teacher judgement data of 1997 and 1998. These are developed independently for tests and then teachers and then the two data summaries compared.