Appendix 6 North West Evaluation Association -Data confirming learning trajectories

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 non-profit organization operating since 1977, which provides assessment products and services to US schools, school districts and states to measure and promote academic student growth. More than 3 million students have been assessed through NWEA, which has established a rich database of student assessments. NWEA use a measurement scale that has been confirmed by regular evaluation to be stable and valid over time (McCall, 2006). The vertical scale is based on the Rasch model. The Rasch model allows the 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.

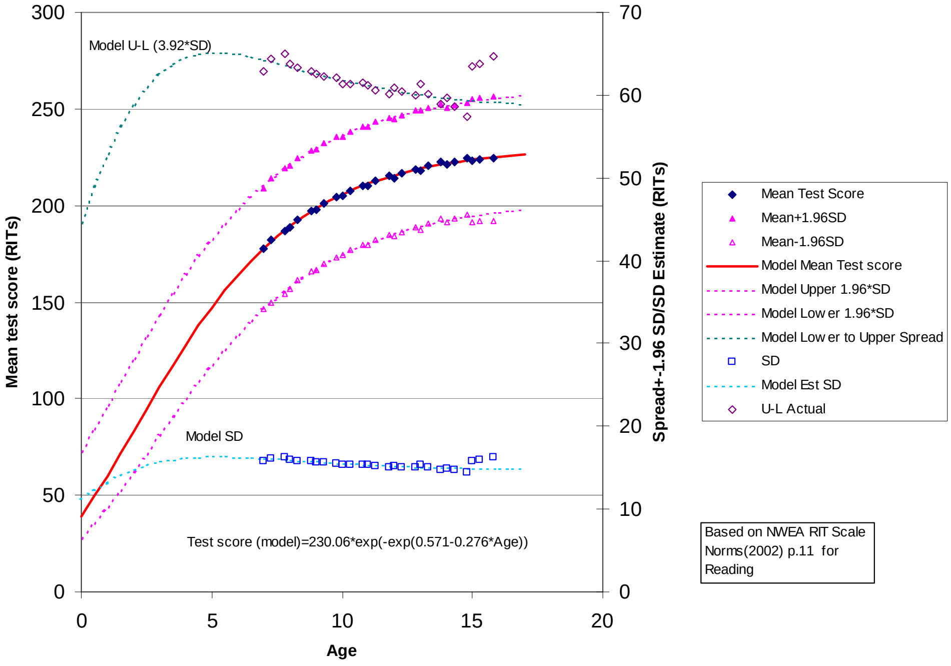
Most of the tests are adaptive and are vertically scaled, drawing on an item bank of 15,000 items. Tests are completed at a computer screen and the process adjusts the difficulty of the items to the current ability of the student. As a result the tests are grade independent. The scale is equal-interval, which allows users and researchers to apply mathematical processes to the scores to establish mean and median scores in a class or grade. The stability of the scale allows individual mapping of leaning growth, as well as valid group comparisons over a span of 20 years of data. (NWEA website, 2009)

The data held are from students over a large number of US states. The data have been used to provide general norms for the typical pathway of development from a range of perspectives. The norming process has established the general patterns of learning growth, the improvement in scores between testing periods, which is related to where on the scale the student is placed at any time.

Growth Trajectory based on NWEA norms

The data from the NWEA norms (2002) indicate a similar pattern to the NAPLAN data, but with more data points at multiple time points for each grade. Figure A6.1 plots the mean test score for each of 9 grades at three points within the grade (Fall, Winter, Spring). The winter data points are interpolated by the NWEA researchers. The time axis is 'estimated average age' of the grade cohort at testing, estimated by the author.

Figure A6.1 NWEA Reading Norms data (2002) with fitted curves.



To spread .3
(Winter), .0

for the initial average age may be in error by 0.1 to 0.3 of a year. All other time points however maintain their correct time relationship to this starting value from this point on. As a result the zero age point is only approximately placed.

A curve is fitted through the points using the Gompertz relation as also applied for the NAPLAN data. (As for NAPLAN a fourth order polynomial also traces the same curve through the actual points but turns downwards after the last data point.) The Gompertz solution was achieved in four iterations and has an asymptote at 230.06 RITs. Following the fitting of the curve the test value at age=0 can be estimated. Since this was positive the original RIT scale was used untransformed. Using the same curve fitting approach, curves are fitted to the upper and lower boundaries for the 95% spread of the data, established by adding and subtracting 1.96 multiplied by the Standard Deviation for each data point. These fitted curves also have positive intercepts on the test scale axis (72 and 27 RITs) and asymptotes at 261 and 203 RITs. The interpolation point is lower than for the NAPLAN model, at about age 3.5 as against age 5.5 for the NAPLAN model. As shown in Appendix 5, the interpolation point can be varied by changing the value of the scale at age zero. In principle all models should assume a common age for the maximum rate of learning. More data are required to establish what this age should be.

Cross-sectional or Longitudinal data sets- do they differ?

Longitudinal data are required to follow the development of individual students and the requirement for individual/personalised data is addressed briefly in the Chapter 5 and in Appendix 10. When the data are summarised as means and SDs do the means differ if the population is large and thus representative?

The NWEA norms (2002 version) are based mainly on cross-sectional summaries rather than longitudinal panels, with grade cohorts ranging from 5000 to 86000 cases, with the mean cohort being over 60,000. A complimentary study (McCall, Hauser, Cronin, Kingsbury & Houser, 2006) examining the trajectories of sub-groups of students to understand the detail of achievement gaps, used longitudinal data obtained from the same data pool. Students from Grade 7 were compared to their position in Grade 4. In this case the total grade cohorts were of the order of 100,000 students. The actual mean scores for the research group and the earlier norming groups above differ at each age, partly because they are calculated on different bases. The cross sectional data had reference time points in fall, winter (interpolated) and spring. The longitudinal data reported the average of 3 to 4 computer adaptive testing sessions at each grade. However the growth between fixed points is approximately the same. The mean scores for Reading in Grades 4 and 7 in the longitudinal study are 198.9 and 214.9 RITs respectively (McCall et al., 2006, p. 17). The fall cross-sectional norms for the same grades are 198.9 and 214.4 RITs (NWEA, 2002, p. 11). The closeness of the values suggests that, in broad terms, the aggregate means of large populations for cross-sectional and longitudinal data are very similar and more importantly, the general growth is similar.