MEASURING GROWTH IN EARLY NUMERACY: CREATION OF INTERVAL SCALES TO MONITOR DEVELOPMENT

MARJ HORNE
AUSTRALIAN CATHOLIC UNIVERSITY

GLENN ROWLEY
MONASH UNIVERSITY

There is a need to measure learning and quantify growth. This is particularly important when looking at any curriculum developments and programs to improve teaching and learning. One off assessments (eg TIMSS) provide a basis for comparisons but do not inform curriculum, teaching or an understanding about the learning process. This paper reports on the development of a scaling process that allowed the assessment to be driven by a model of learning and mathematical development rather than the statistical process. The process then allowed the scales to be transformed to interval scales enabling growth to be investigated and comparisons to be made between groups.

When the Early Numeracy Research Project (ENRP) began at the start of 1999 there was a need to develop a comprehensive and appropriate learning and assessment framework for early numeracy. Drawing upon Australian and overseas research on young children's mathematics learning, the ENRP team developed a framework of the key "Growth Points" in the early learning of mathematics. For teachers these growth points, from a number of mathematics curriculum areas, can be used to describe children's numeracy development and to inform their teaching. For example, those in the Number area include Counting, Place Value, Addition and Subtraction, and Multiplication and Division. The process of development of the framework and its associated assessment interview is described more fully elsewhere (Clarke, Sullivan, Cheeseman & Clarke, 2000; Clarke, Gervasoni & Sullivan, 2000).

The purpose of the framework was to inform teachers in their planning and implementation of curriculum and to provide a basis for assessment which would enable children's mathematical development to be monitored. The assessment was to provide information for the research team on the development of aspects of numeracy, but it was also to inform teachers, enabling them to learn more about the thinking and development of individual children in their classes and to plan their teaching accordingly. Assessment, in the form of a 30-40 minute task-based interview, was developed based on this framework. The nature of this assessment is essentially different to the one off standardised assessments.

- It is developed from a structure of mathematical development rather than from a content base.
- It does not require all children to answer all the same questions but is rather structured so that students stop answering questions within a domain once there is a lack of success.

- It allows for a focus on the approaches the child uses as well as the answers given.
- Its purpose is to inform teaching as well as to provide information on children's mathematical development.

Data collected from interviews with over 5000 children in each of March and November in 1999 and 2000 have led to minor modifications of the framework and enabled monitoring of some aspects of children's mathematical development. To date, 4-6 Growth Points have been developed for each domain. It is possible that more Growth Points could be added, as the research and the interview protocols are extended to higher grade levels.

The ENRP is a large scale project looking at a whole school approach to school improvement in early numeracy. It is a developmental project with a team of researchers working with all teachers of early numeracy in the 35 trial schools. One aspect of investigating the development of early numeracy is to compare growth between students in the 35 trial schools which have been chosen to be representative of region, size and socioeconomic indicators and the matched 35 reference schools, at each of the grade levels involved. Another purpose is to study growth and inform teachers of the way children learn and their numeracy development. Such comparisons are obviously facilitated by the use of an interval scale of measurement. While the growth points were chosen to be important ideas and are not exhaustive but rather a set of sign posts, they were not chosen with any interval properties in mind.

One approach which has been used to look at underlying scales with interval properties in recent years has been Rasch modelling (Andrich, 1988; Pirolli & Wilson, 1998). The data from the interview are not the equivalent of students answering items on a written test as the interview within a domain ceases when errors are made so the children do not all answer the same set of questions, and, although some questions may be common, the full range is not appropriately sampled. Although Rasch models do allow for missing data, the cutting off of the questions once lack of success is experienced makes the use of Rasch modelling problematic.

The study of growth with such a large number of children and many domains would be enhanced by the use of an interval scale. Without an interval scale, growth differences between children are only clear when the starting points are the same. An interval scale would enable more appropriate analysis of such a large amount of data. This paper looks specifically at a new approach to the development of interval scales and reports some findings based on those scales.

Analysis

A large sample of students has been involved in the two years of the project – to date 8681 students from 68 schools balanced with respect to region, size and socioeconomic indicators. Not surprisingly, the frequencies of responses in each domain of the interview strongly suggest that the growth points do not form an

interval scale. Figure 1 shows the distribution of Growth Points for the domain *Counting* in two successive years. The distributions are irregular, but consistent from year to year. The uneven distribution suggests that Growth Point 2 spans a wider range of performance than Growth Points 1 and 3 on either side. Similar patterns occurred with other Growth Points. The stability of these patterns over successive years indicates that they reflect a property of the measurement scale, rather than arising from peculiarities of the samples. Essentially, the implication is that the student development from Growth Point 1 to 2 may not be the same as that from 2 to 3, etc. While this is generally true of achievement test scores, it inhibits making comparisons between the growth of one group and the growth of another, unless they start at the same point. A key purpose of the ENRP is to demonstrate growth to teachers in a way that is meaningful to them, and to find how this growth is enhanced by the whole school approach and professional development that the project offers. For this reason, we have investigated adjusting the location of the Growth Points along a scale of achievement so that they represent achievement within a domain on a scale with interval properties.

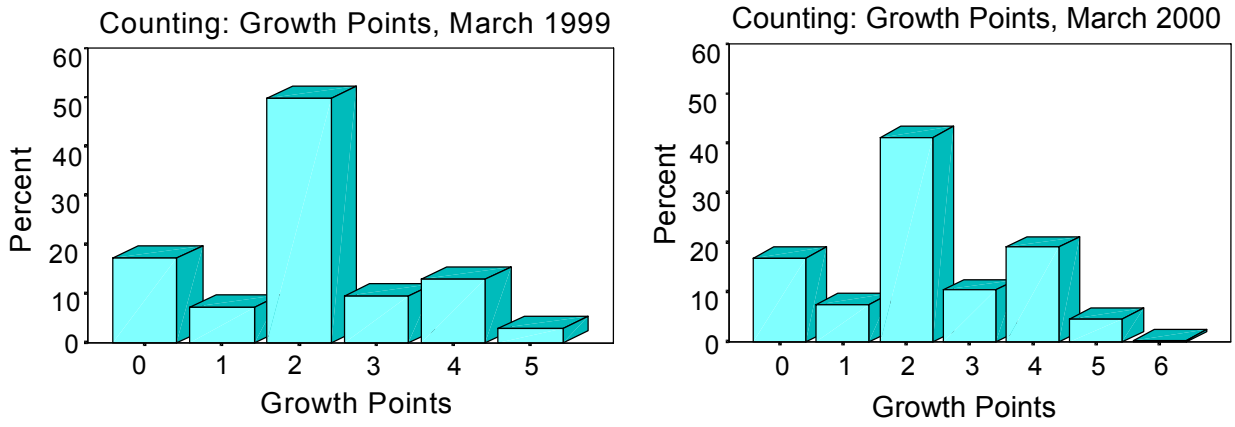

Figure 1. Distribution of growth points for the counting domain in March 1999 and March 2000.

Creating an interval scale

The major assumption that is made in the scaling process that follows is

> *that the nature of the distribution of learning in the population of children in the first three years of school is normal.*

The measurement of other attributes of a person such as height, with a cohort like this one, do approximate a normal distribution. As long as there is a continuum of learning in the domain, beginning before children commence school and extending beyond grade 2, the spread of children within the domain should approximate to a normal distribution. We can then regard the Growth Points as indicators of an underlying normal distribution, cut into slices of unequal width. The sample of students used is wide enough and representative enough that, given the above assumption, there is an underlying normal distribution of performance. It is not uncommon for researchers to assume a normal distribution for a measure, even when

there is no particular reason to expect it. We emphasise that in this case, the assumption is not made lightly. The justification for this is that the sample is large, and that the schools have been selected to be representative with respect to geographic area, school size, and socioeconomic indicators. The test is whether scaling from two different populations yields similar sets of scaling points.

This adjustment of the location of the Growth Points along a scale was done by standardising the scores, then re-scaling them using the mean and standard deviation of the original set of Growth Points. With this process it is reasonable to expect similar scaling points from populations that differ in mean and standard deviation. In detail the steps were

1. We chose the median student to be representative of the growth point then calculated the proportion of students below, thus giving a cumulative frequency distribution. This means that below the median student were half the students on that growth point plus all students on lower growth points.
2. Assuming a normal distribution, we found the z-score of the median student for each growth point by using the probability from the relative cumulative frequency. The probability of being below the median student in a growth point is the number of students below (calculated in step 1) divided by the total number of students. This was then looked up on a normal distribution table to give a z-score for the median student representative of the growth point.
3. We translated this z-score back to a distribution with the same mean and standard deviation as the original data.

The Trial and Reference Group data from the March 1999 interviews were rescaled in this way. These are two large populations (n = 3639 and 1219, respectively) representing a wide range of geographic regions and socioeconomic characteristics. Table 1 shows the transformations, as applied to these two populations.

Table 1: *Scaling of growth points for Counting: trial and reference schools, March 1999*[a]

	Trial Schools		Reference Schools	
	N	**Scaled Growth Point**	**N**	**Scaled Growth Point**
0	588	0.287	254	0.256
1	242	0.984	112	1.002
2	1802	2.016	612	1.989
3	365	3.064	103	3.094
4	527	3.721	107	3.678
5	115	4.871	31	4.620

[a] Rows are shaded for n < 50. With so few people in the score range, the accuracy of the scaling may be problematic.

Comparison of the scaling from the two distinct populations indicates that it matters little which population is used to scale the Growth Points. For example looking at the scaled growth points for growth point number 2 in table 1 shows the scaled score for the trail schools to be 2.016, only just above the hypothetical growth point 2, while for the reference schools the scaled growth point at 1.989 was only just below. The

difference between the two independent populations for this growth point was only .027. For 5 of the 6 Growth Points shown here in Table 3, the difference is less than 0.10 in magnitude, and the remaining one arises from a Growth Point with small n (less than 50, shaded in Table 3.) and has a difference of about 0.25. We assert that the process we have adopted yields results that are relatively invariant across populations, provided there are sufficient data at each Growth Point.

The scaling process yields scores that can be mapped onto scales with any size units. Given that teachers and others may still wish to interpret growth information and mean scores in terms of the Growth Point scales with which they are familiar, a case can be made that the most suitable mapping is onto a scale in which the two endpoints remain fixed, and the intermediate points are adjusted. Table 2 shows the results of such a scaling in the column labelled Scaled Growth point 0-6 Scale, using the entire March 2000 cohort, including the grade 3 and 4 children, (Trial and Reference schools) as the scaling population. The process followed was the same as the one above to give the growth points in the column headed Scaled Growth Point but with a fourth step included that stretched the scale from 0 to 6 while maintaining its interval nature. Again looking at the hypothetical growth point 2 in Table 2 the scaled growth point obtained by using the original process but the different population was just above at 2.08. One the re-scaling has been done to set the lower and upper growth points at 0 and 6 this scaled score becomes 1.83, just below the 2.

Table 2: *Scaling of growth points for Counting: trial and reference schools, March 2000*

	N	Scaled Growth Point	Scaled Growth Point 0-6 Scale.
0	983	0.24	0.00
1	444	1.07	0.82
2	2440	2.08	1.83
3	642	3.02	2.76
4	1320	3.74	3.47
5	492	4.92	4.65
6	69	6.28	6.00

Use of the scaled Growth Points

One of the ways in which the research data might be used is to compare groups. With some confidence in the interval properties of the re-scaled Growth Points, we are now in a position to reach conclusions about relative growth, even from comparison groups that have different starting points. Figure 2 shows growth on the *Counting* domain. The growth over the 1999 and 2000 academic years is apparent, as is the slower growth over the 1999-2000 summer break. Comparisons across cohorts show the gap between grade levels. In each case the graphs are in pairs showing trial and reference schools in a particular cohort. For example the upper two graphs on the left are the cohort of students that began in grade 2 in 1999 and in 2000 were in grade 3 and thus were not tested as they were no longer part of the project (a small sample was tested in 2000 but not as part of the main project). The two lines below

these represent the cohort of students who began in the project as grade 1 students in March 1999 and were tested in Nov 1999 and in March and November of 2000. It can be seen that in the trial and reference schools grade 1 students were very similar at the beginning of 1999 but that the growth in the trial schools was greater during both 1999 and 2000.

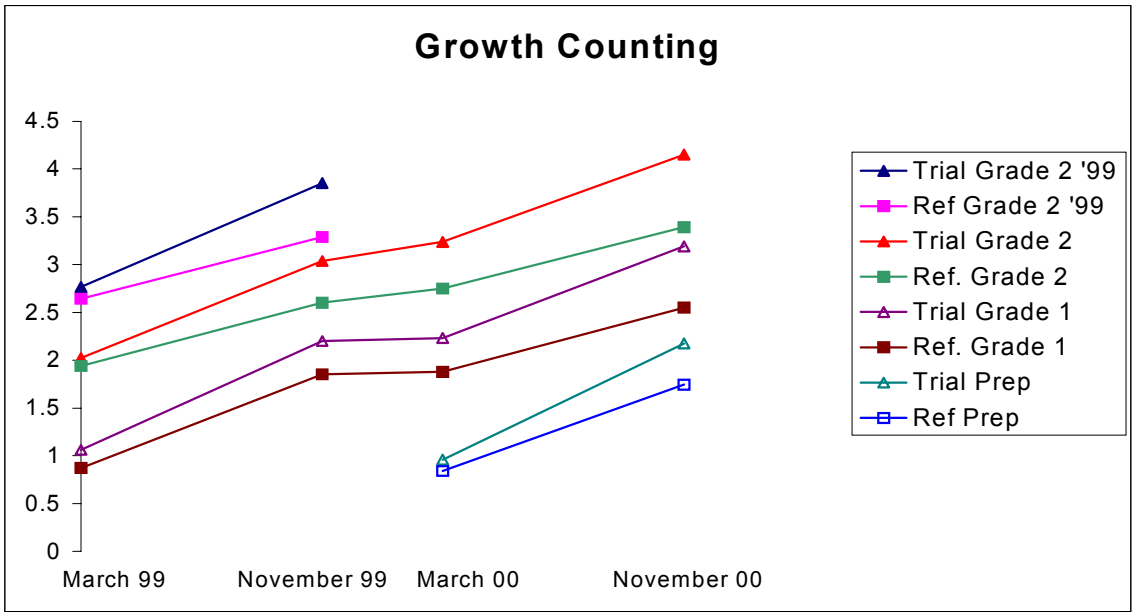

Figure 2: Growth in *Counting*, 1999-2000 in trial and reference schools

The interval scales can now be used to compare growth for investigating successful aspects of the program and to inform teachers of growth for their students and classes. The scales have enabled investigations such as the growth is typical in a year and the extra growth that occured as a result of the ENRP program provided in schools. Teachers involved in ENRP understand a scale of measurement framed in terms of Growth Points, and welcome information about achievement in this form. Until we were confident that Growth Points could be manipulated into an interval scale, we were unable to provide this information in a way with which we could feel comfortable. Because the transformation described previously is fairly close to the unscaled Growth Points, we can present information about means and variances in terms of scaled Growth Points, and remain confident that this information will not mislead users who interpret them in terms of understanding and skills possessed by their students. Because they are close this also means that we can refer to the growth points as whole numbers knowing that they are close to the scaled values. In terms that are readily understandable by teachers in the project, we can say that typical gains for students in Reference Schools are a little less than one Growth Point per year in *Counting*, but a little more than one Growth Point per year for students in Trial Schools. Very little of this gain occurs between November and March, a good part of which is summer holidays.

The differences found between trial and reference schools are statistically significant. Rowley and Horne (2000) reported an analyses of variance using the scaled March

achievement as the covariate with the independent variables of schoolcode, grade level and ENRP participation (i.e. Trial versus Reference Schools). For all six domains used in both years, they found a significant ENRP main effect, indicating that the effect of ENRP on achievement can be detected, over and above that of prior achievement, school attended and grade level. A repeat analysis done with the year 2000 data supported the same findings.

Finally, the interval properties of the scaled Growth Points allowed us to test the applicability of particular models to the data. In this example, the models included, separately at each grade level, in logical-temporal order:

- The school that the child attended (as represented by the matched pair of schools, one Trial and one Reference school)
- The child's prior (March 1999) achievement
- The provision of ENRP Professional Development in the school.

Regression analysis allows us to ask whether each of these in turn produces a significantly better-fitting model than those that precede it. In essence, this tells us how well is achievement in November predicted by the school attended, achievement in March, and by whether or not the school participates in the ENRP Professional Development program. And finally, we can also see how much difference does the ENRP Professional Development makes by using the re-scaled growth points.

Table 3: *Regression coefficients for provision of professional development for Counting*

	1999	**2000**
Preps	0.256	0.385
Grade 1	0.376	0.381
Grade 2	0.466	0.481

Table 3 shows the unstandardised regression coefficients for ENRP Participation in the full model, which includes School, Prior Achievement, and ENRP Participation. These coefficients may be interpreted with some precision as the expected advantage in achievement, assessed in Growth Points, that students gain through their school's participation in the ENRP program. In general, students may expect to gain from 25 percent to 48 percent of a growth point per year more if their school participates in ENRP program.

Conclusions

In this paper, we have presented the construction of interval scales based on a set of protocols for one-to-one interviews of children in their first three years of schooling that can be used to assess their development on curriculum-relevant domains of mathematical learning. Because of the nature of interval scales, the assessments obtained are useful for assessing and charting growth.

Using the Growth Point Scales in this way, we have shown some representations of growth in achievement, and demonstrated a modest but consistent advantage in achievement from participation in the ENRP program. This rescaling process has provided data which has been used for comparisons to investigate differences between groups looking at the success of programs and, perhaps more importantly, to look at children's development in early years numeracy.

The scales for separate domains are referenced to Growth Points within each domain, but not to each other. Future use of the data will enable typical growth in the domains to be investigated. At the start of grade 2 before the project began, for example, students were typically at Growth Point of 3 in *Counting*, but just at growth point 2 in *Multilpication and Division*. Equally, a typical year's growth is approximately one Growth Point in *Counting*, but only half that in most other domains. While these comparisons are possible we also do not want to lose sight of the many other developments which are not measured by these interview based scales. The comparisons now possible are being used as one aspect of deciding on case studies to be carried out in 2001.

Acknowledgements

The main team of researchers involved in the ENRP includes D. M. Clarke (director), J. Cheeseman, B. Clarke, A. Gervasoni, D. Gronn, M. Horne, A.. McDonough, P. Montgomery, G. Rowley, and P. Sullivan. All of this team have been involved in the research presented here.

References

Andrich, D. (1988). *Rasch models for measurement*. Sage university paper series on Quantitative applications in the social sciences, 68. Newbury Park, Ca: Sage Publications.

Clarke, D. M., Sullivan, P., Cheeseman, J., & Clarke, B. (2000). The early numeracy research project: Developing a framework for describing early numeracy learning. In J. Bana & A. Chapman (Eds.) *Mathematics Education Beyond 2000*. (Volume 1 pp 180 – 187) Perth, WA: Mathematics Education Research Group of Australasia Incorporated.

Clarke, D.M., Gervasoni, A. & Sullivan, P. (2000). *The Early Numeracy Research Project: Understanding, Assessing and Developing Young Children's mathematical strategies.* Paper presented at the Australian Association of Research in Education conference, Sydney, Dec, 2000.

Pirolli, P. & Wilson, M. (1998). A theory of the measurement of knowledge, content, access, and learning. *Psychological Review* 105 (1). 58-82.

Rowley, G. & Horne, M. (2000). *Validation of an interview schedule for identifying growth points in early numeracy*. Paper presented at the Australian Association of Research in Education conference, Sydney, Dec, 2000.