Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE   Sydney,   BON05361   1 of 16

BON05361

Silk purses from sows' ears? Making measures from teacher judgements[1]

Trevor Bond, Hong Kong Institute of Education (tbond@ied.edu.hk) and
Martin Caust, James Cook University (mkcaust@fastmail.fm)

The prospect of increased mandated achievement testing in Australian schools has the potential to relegate into insignificance the professional judgements of classroom teachers. This paper reports the first steps of a two-stage project to foreground teachers' judgements in the assessment process. Firstly, we investigate the extent to which teachers' assessments of their students satisfy the strict measurement requirements of the Rasch model. Secondly, we attempt to integrate the ability estimates derived from teacher judgements into the more typical quantitative results derived from standardized testing. Two data sets (from 1997 and 1998) record teacher assessments of the development of approximately 10,000 primary school students for each calendar year using the Australian National Profiles (reported by Rothman, AARE, 1998). As a separate unrelated event all students in Years 3 and 5 were also assessed using a Literacy scale using the NSW Basic Skills test. With the recent approval and support of the SA school system, the 'intersect' of the two sets has been matched at Years 3 and 5, in English in 1997 and Mathematics in 1998. (1000 teacher assessments at each Year level matched with 12,000 test assessments). Files of approximately 700 students with both assessments have been created for the 4 data sets to explore (a) the development of a measurement scale of teacher assessments and (b) how well the two approaches to assessment of students match. Measurement error based on the Rasch model has been estimated for both the test and for teacher judgements. Teachers vary considerably in their observational skills, their understanding of learning, their comfort with the ambiguous profile scales, their personal specific knowledge of the randomly selected students and their confidence that they can use criteria described scales. Many teachers' assessments correlate well with the test, some differ widely. The paper then speculates on how to improve the skill of teachers in using the latent scales established in test analysis as a support to the measurement of growth in the classroom and, more generally, how to use criteria scales in formative assessment.

Keywords: Assessment and measurement

Background

Current rhetoric about educational measurement and assessment policy encourages the measurement of developmental growth of individual students (Masters, 2004; Griffin, 2004; Kingsbury, 2000; Cronin et al, 2005; Hauser, 2003; Wilson, 2004) among others. A variety of views are expressed about how growth might be made the focus of teacher and school activity, with some advocates encouraging regular and increased testing (NCLB Act). The US Department of Education has agreed (November, 2005) to allow 10 states to explore individual student growth approaches to meet the requirements of the NCLB Act (Associated Press, 2005).

At the teacher/classroom level some commentators point to the poor quality of classroom assessment, as far from reaching "*or even approximating its immense potential as a school improvement tool*" (Stiggins, 2001). Classroom assessment, as a critical tool to improve the learning development of all students, is encouraged by Brookhart (2003, 2004). Marzano (2000) advocates significant improvements in classroom grading practices for American schools and identifies methods of monitoring student progress as one of the keys to improved student learning (Marzano, 2003).

Masters (2004) advocates '*measures of progress made by all students in a grade*' while accepting that this need not '*replace information about the percentage of students meeting grade level expectations*'. He argues that improvement in learning '*depends on an understanding of the variation in students' level of development within the same grade*' as well as a '*willingness to monitor and report individual growth*' across year levels for any student. He proposes graphing of students progress trajectories within and across grades,

---

Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   2 of 16

drawing data from Rasch calibrated measures such as research-based progress maps, many of which draw on teacher observation and judgement and do not necessarily require testing. He makes the key point that the ubiquitous letter grades (A,B,C,D,E) are inadequate for monitoring and reporting growth across the years of school since many students obtain the same grade, be it A or D, year after year, giving the distinct impression that no progress has been made. He reiterates that grades are also unable to describe what a student has actually learnt nor what s/he is capable of learning.

Griffin (2004), drawing on the work of Rasch, Glaser and Vygotsky argues for strategies that help keep the zone of proximal development (ZPD) for each student in the forefront. To help achieve this he advocates better appreciation of the probabilistic nature of statements about student performance, the integration of item response modelling, criterion-referenced descriptions of student learning and the ZPD. He proposes that teachers can apply these '*approaches to teaching, assessment and learning in their classrooms without large scale adoption of sophisticated computer models*'. He also advocates mapping and graphing of student progress.

The US Northwest Evaluation Association series of regular papers (Kingsbury, 2000; 2004; Cronin et al, 2005; Hauser, 2003) has explored aspects of student growth drawing data from its longitudinal data base to illustrate rates and spreads of student achievement over multiple grades and time. The key to the NWEA process has been the maintenance of a large database of student performance on common Rasch modelled scales.

A number of education systems and commentators have emphasized the value and priority of teacher judgments in classroom assessment (South Australia, Victoria, Queensland, Maine, Nebraska). Guskey (1996) asserts that teachers '*know their students, understand various dimensions of students' work, and have clear notions of progress made.*' As a result '*their subjective perceptions may yield very accurate descriptions of what students have learned.*' Marzano (1998) encourages observation as an unobtrusive way to assess students' competency as they go about their daily business. Observations use teacher judgement to place students' performance '*somewhere on a continuum of achievement levels ranging from very low to very high*'. Shepard (quoted in National Research Council, 2003) asserts "*The best way to help policy makers understand the limitations of an external, once-per-year test for instruction is to recognize that good teachers should already know so much about their students that they could fill out the test booklet for them.*"

Teacher judgement requires teachers to integrate data about students from many sources in a way that is manageable.   We need to develop methods of observing and recording student development provide regular assessment data values over time, without necessarily requiring 'high-cost high-tech' support. One option, that has not been fully explored, even from the point of view of data available, has been the Australian development of *The Statements and Profiles for Australian Schools (1994)* (SPFAS). Descriptions of criteria to be met were developed at 8 levels in 8 learning areas, with each learning area made up of a set of strands. Each strand (Reading and Viewing as an example) had descriptions for 8 levels of achievement.

Assuming the distance between levels could have been refined overtime to sort out any anomalies, as applied to a music curriculum (Bond and Bond, 2003), the Profiles could have been iterated to be both a general curriculum guide as well as a learning development metric. Data from one school system (South Australia) that used SPFAS as a data collection framework for 2 years, are being re-analysed to explore the link between teachers' 'on-balance' assessments and mandated test assessments.   This analysis is based on a key assumption, that Literacy (a test dimension) and the English (Reading, Writing and Speaking strands) Profile levels (a curriculum dimension) could be assumed to be approximately the same dimensions.

The object of this paper is to present a preliminary analysis of historical data collected in 1997, two parts of a four part data set for 1997 and 1998. The data combine teacher and test assessments for a common set of students and then extends the analysis of Teacher assessments over 8 year levels. We explore the accuracy of teacher 'on-balance' judgment of

Paper Presented at the AARE Annual Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   3 of 16

students' developmental position on a latent dimension relative to an independent measure, a test assumed to be placing students on the same scale. The purpose is to provide data, albeit historical, to contribute to the debate on appropriate methods for obtaining regular data points in order to monitor student learning growth. Insights into processes that maximize the use of teachers as the creators of these data points and that also meet the multi-benefits of cost-effectiveness, enhanced pedagogical focus on individual student development and enhanced professional confidence in and by teachers are possible from this data set.

There is already a moderate literature on the ability of teachers to be judges of student development and achievement. In this context 'judgement' implies a process of drawing together relatively quickly (usually in the mind without recourse to extensive external data collation), impressions based on conscious and unconscious observations and recollection of students' behaviours, questions asked by students, test results, portfolios, observed misconceptions to form an 'on balance' judgement of student learning. Often what is missing in such a judgement is adequate training or preparation in the metric to be used to describe or articulate (i.e. locate) the judgement. The assumed latent dimensions for Profiles for Australian Schools might have been the beginning of a metric, no doubt requiring significant refinement, to describe and record teacher judgements efficiently. Such a metric would have value if the curriculum, external testing, classroom assessments and teacher-teacher and teacher-student-parent dialogues were then conducted in a common language. This new language might eventually replace the entrenched, archaic and poor-measurement-properties of grades, percentages and ranks still applying in many classrooms.

Improved formative assessment practices as advocated by Black and his colleagues (1998, 2004) require formal and informal methods to keep teachers and students informed about where each student is starting and whether learning and understanding are occurring. An improved short-hand for teachers to note each student's position and development could be developed from the ideas considered in this paper.

A variety of metrics have been used in teacher judgement investigations. Meisels et al (2001) applied the metric of existing test batteries. Fuller (2000) explored teacher ability to predict the likelihood of students passing 4th and 6th Grade proficiency tests. Coladarci (1986) tested Shepard's assertion above (though many years before she expressed it) and found teachers were able to estimate how students would complete a test paper (SRA Achievement Series, Science Research Associates, Inc., 1978). Correlations of between 0.62 and 0.98 were obtained on an item per item basis between the teacher and the test. Bates and Nettelbeck (2001) explored the ability of teachers to predict reading achievement among the same general population of teachers that is included in the analysis in this paper, South Australian primary teachers circa 1997. Among other findings the authors concluded that teachers tended to '*over-estimate the relative percentile position of children performing less well and under-estimate the achievement of better readers*'.

Both the Victorian and South Australian school systems have encouraged teacher 'on-balance' developmental assessment and mandated, for periods, estimating the position of students on scales based on to Profiles for Australian Schools (SA), or closely related to them (Victoria). Using the Curriculum Standards Frameworks (CSF 1 and 11), Victorian teachers have used a three-zone interval to report where in the zone between adjacent levels a student performance lies (Office of Review, Victoria, Benchmarks 97, 1998). Teachers indicate a zone for each student based on their judgements of student mastery of required skills using 'beginning', 'consolidating' and 'established' to signify the location.

On this basis some students, echoing the inadequacy of grades described above by Masters, could go for two or more reporting periods without showing a change of zone. Benchmarks (i.e. state means), based on this approach had a mean difference from one year level to the next of about 0.45 of a level in 1997 (Office of Review, 1996 & 1997). The assessment 'zones' are about 0.33 of a level, equivalent to about eight to nine months of development, on average. Results of the annual state-wide test for grades 3 and 5, the Assessment Improvement Monitoring (AIM) test, are reported on the same 7 level scale as teachers have been required to use, broadly, the same scale for reporting to parents. Research that analyses

Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   4 of 16

the matching of test and teacher assessments at an individual student level using the CSF scale might exist but has not been found. Victoria however is one of few education systems where tests and some classroom assessments are reported on the same scale over an extended period. It appears to be a unique system where teachers are adequately familiar with common scales for useful further research on teacher's ability to judge on the test scale to be conducted.

For the years 1997 and 1998, South Australia required the collection of student assessments from a large sample of teachers (Profiles Collections, 1998 & 1999). Data was collected from Year 1 to Year 7 teachers in primary schools and Year 8 teachers in secondary schools (estimated at 1600 teachers in 1997, 3000 in 1998 of a possible 10,000 teachers in these year levels) with a sample of 4-6 students, randomly sampled by computer, for each teacher. On the insistence of the Australian Education Union (SA) two elements were added to the collection. Nine progress zones covered the 'territory' within a level (as distinct from just the three zones above for Victoria) and secondly teachers had to indicate their confidence in the general process and their confidence in their assessment for each sampled student. The reasons given by the union were

- to be able to show student progress from one collection to the next; and
- to be able to show changes in confidence in the process and to identify when less confident judgments for some students were made (say, when a teacher had only limited contact with a student.)

The progress points between levels were estimated on the basis of the teacher's perception of the proportion of criteria for the next level already met. Charts of the median and spread for each learning area, presented by year (grade) level matched the data published elsewhere in Australia (Office of Review, Victoria, 1997), (Rowe & Hill, 1996) and are described extensively in Rothman (1998, 1999).

Data from the 1997 and 1998 South Australian collections have generously been made available, along with student test data at years 3 and 5 for 1997, and 1998, with keys to enable the matching of some data at an individual student level. This paper covers the progress of an initial analysis of the Year 3 and 5 data from 1997 and explores the extent to which teacher and test assessments appear to match.

The Data

In 1997 and 1998 the then South Australian Department of Education and Training tested students at Year 3 and Year 5 in Literacy and Numeracy, using the NSW developed test of Aspects of Literacy and Numeracy, commonly known as the Basic Skills Test. In these two calendar years the Department also collected data from schools, for a randomly selected sample of students, based on the 8 Learning Areas of the Statements and Profiles for Australian Schools (Curriculum Corporation 1994) as described above. In 1997 the four Learning Areas collected included English. This teacher assessment data were collected in October 1997; the students had been tested in August. The student-based teacher assessment collection process lasted for the years 1997 and 1998 only as the SA Department commenced a revision of its curriculum framework, leading to the implementation of the SA Curriculum and Assessment Framework, with different emphases and structures relative to the National Profiles.

Test data at Year 3 covered, approximately, the full population of Year 3 students. Data were provided for 12437 cases. Test Data for Year 5 were again notionally the full government school population, and covered 11,972 cases. The 1997 Tests have already been part of an extensive publicly reported longitudinal analysis (Hungi, 2003).

Teacher assessments, restructured to consolidate the three assessments for each student, covered Reading and Viewing, Writing and Speaking and Listening. 7872 students from Years 1 to 8, approximately 1000 per year level, are included. The Teacher collection for 1997 created 130,000 records in the form of a case of a teacher assessment of a student in a

Paper Presented at the AARE Annual Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   5 of 16

strand of a learning area. These cases covered year levels 1 to 8 and have been well summarised by Rothman (1998; 1999). All cases were anonymous and no identifying teacher information was ever collected. The collection process created a unique identifier for each student for collection management purposes. This identifier has, 8 years after the collection, enabled with considerable effort, the matching of Test and Teacher assessments for a part of the sample of students.
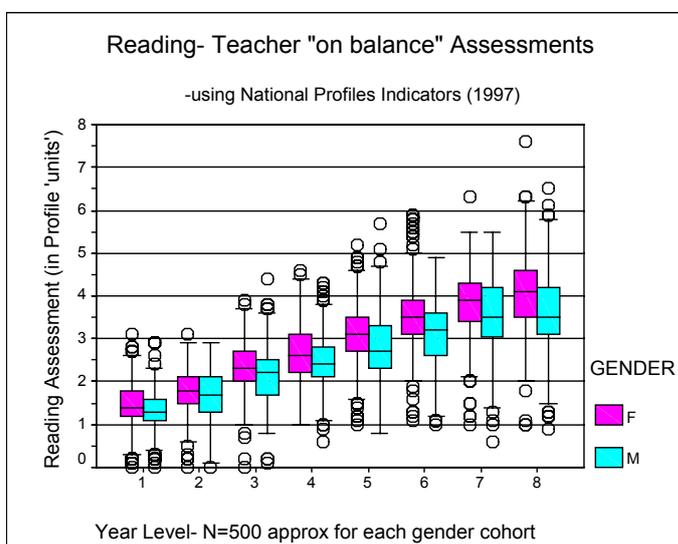
Through a three stage matching process 1275 test cases at Years 3 and 5 combined were matched, from a possible 2002 (1006 + 996) Teacher Assessments at Years 3 and 5, i.e. approximately 5% of Test cases were matched to about 60% of the Teacher Assessments at these two year levels. The teacher assessments were of strictly randomly sampled students and it is assumed the matching process was unbiased.

The Teacher assessments were made in a two-stage process. Criteria to be met to achieve any particular Level in a strand were provided to teachers (and are described in Statements and Profiles for Australian Schools (Curriculum Corporation, 1994)). It had been assumed that most teachers were familiar with the descriptions relevant for the range of students whom they were teaching, particularly in the English Learning Area. A concerted teacher development campaign had been in place for the three years leading up to the collection. This campaign was consistent with a previous teacher development campaign to help teachers understand curriculum assessment described in 'achievement levels' that had run since 1989. However, while many teachers were very familiar with the criteria levels, the criteria descriptions were still made available to the teachers during the assessment exercise. For each strand for each student the teacher had to indicate which of the levels the student had most recently fully achieved. In addition the teacher had to indicate an assessment of the student's progress towards achieving the criteria for the next level. Effectively it divided the distance between two 'major tick marks' on the 'scale' into ten minor 'tick marks' on the basis of teacher-perceived progress made.

Documenting the assessments was computer-assisted, with the teacher responding to a computer presented proforma, which had already automatically sampled the students. The teacher entered the level (a number) and then indicated progress towards the next level by clicking at a point on a horizontal bar. This bar was divided into 9 hidden segments, effectively dividing the progress into 9 decimal points. The data value for the student for a give assessment had two parts, an integer value and a progress value. For example a student who met the criteria for level 2, and was deemed to have met about 0.3 of the criteria for level 3 was reported as 2.3.

One form of the original analysis was to show all cases by Year level in a box plot. Figure 1 shows the pattern and spread with each year level for the Reading Strand.

Figure 1 Teacher Assessments of Reading, 1997

Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   6 of 16

The graph of teacher judgements is based on 8000 cases, approximately 1000 per year level. The data reflect patterns and spreads comparable to other studies (Rowe & Hill, 1996) using similar teacher assessment approaches, or as shown by Hauser (2003) using test data.

Method of Analysis

General Process

Through a three-stage process, Test cases for 1997 were linked to Teacher Assessed cases at Years 3 and 5. Out of 24, 400 Test cases at Years 3 and 5, 1275 were matched from a possible 2002 (1006 plus 996) Teacher Assessments at Years 3 and 5, i.e. approximately 5% of Tests matched to about 60% of the Teacher Assessments at these two year levels.

All Teacher Assessments (7872, Years 1 to 8) were analysed in Winsteps using three items, Reading and Viewing, Writing and Speaking and Listening, treating the Teachers together as a single instrument with three items in the default Partial Credit Model. The assumption that the Teachers together can be treated as a single instrument is problematic. In practice this assumption leads to reasonable measures consistent with the Rasch model, except the degrees of fit, particularly overfitting data (data that is too good to be true). The resulting measures however correlate strongly (r=.99) with other methods e.g. equi-percentile equating of Teacher units (in Profile units) to the distribution of Test scores in logits. The advantage of the Winsteps process is the ability to estimate measurement error within the model for each case. Assessment values for each item ranged from 0 to about 80 (8.0 on the Profiles scale), that is a Teacher assessment of say 2.3 was coded as 23. The default Partial Credit Model was applied taking advantage of the flexibility in data format allowed by Winsteps, the ability to use two ascii characters to denote a value for each item.

Measures for each student common to the Test and the Teacher data sets were extracted. Teacher measures were rescaled to ensure the Mean and the SD matched the Mean and SD of the Test data to faciltate common person linking. This set of 1275 cases was then explored for the degree of match of the two data sets, taking account both of the location and the errors of measurement established in the Test and Teacher Assessment Winsteps analyses.

The mean difference between the original and rescaled teacher assessments in the step described above was used to rescale all 7872 Teacher Assessments to a scale that approximated the Test scale. This second data set was used to examine other ways of exploring the link between Teacher and Test Assessments.

Results

Table 1 lists the key Winsteps statistics for the three key data sets.

Table 1: Summary of Winsteps Fit and Measurement Statistics

Items	N	Measure	Error	SD of Measure	SD of Error	Reli-ability	Separ-ation	Real RMSE	Adjust-ed SD	Infit MS	SD of Infit	Outfit MS	SD of Outfit
Test Y3	58	0.00	0.02	0.00	0.02	1.00	40.80	0.02	1.00	0.99	0.12	0.98	0.24
Test Y5	83	0.00	0.03	0.01	0.03	1.00	45.11	0.03	1.25	0.99	0.11	0.96	0.23
Teachers	3	0.00	0.00	0.00	0.00	0.92	3.47	0.00	0.02	0.94	0.19	0.93	0.19

Persons	N	Measure	Error	SD of Measure	SD of Error	Reli-ability	Separ-ation	Real RMSE	Adjust-ed SD	Infit MS	SD of Infit	Outfit MS	SD of Outfit
Test Y3	12437	1.03	0.37	1.30	0.11	0.91	3.24	0.38	1.24	1.00	0.12	0.98	0.33
Test Y5	11972	1.42	0.33	1.22	0.08	0.92	3.45	0.34	1.17	0.99	0.15	0.96	0.37
Teachers	7872	-1.47	0.30	1.99	0.13	0.97	6.00	0.33	1.96	0.86	1.54	0.86	1.54

Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   7 of 16

Persons	Above 1.3 Infit MS	Below 0.7 Infit MS
Test Y3	1.8%	0.1%
Test Y5	2.9%	0.7%
Teachers	20%	69%

In all three sets the scales are non-anchored.   Original values for Year 3 and Year 5 measures, already anchored by the original analysts, were used in the matched data sets. The re-runs of the original Test data were used to re-calculate additional statistics not provided in the data released from the SA Department of Education and Children's Services. The table illustrates greater variability some person statistics in the Teacher data, particularly Infit and Outfit mean squares and the very large numbers of Infit mean square cases above 1.3 and below 0.7. The model does however provide estimates of Teacher measures consistent with other methods explored for equating (equi-percentile equating) and in addition, provides estimates of measurement error based on the Rasch model.

How well do Teacher and Test assessments compare?
Figure 2 shows the scatterplot of the 1275 common cases, with the Teacher assessments converted to match the mean and standard deviation of the Test person distribution. Cases are deemed to be equivalent if the Test and Teacher assessment pair are within the 95% quality control zones obtained from the combination of measurement errors for the Test and the Teacher. If the two assessment processes were assessing on the same latent dimension, the proportion of cases within the control lines would be expected to be 95%. In this analysis only 57% of the cases fall in this zone.

**Figure 2 Teacher and Test Assessments, 1997, Cases deemed identical.**
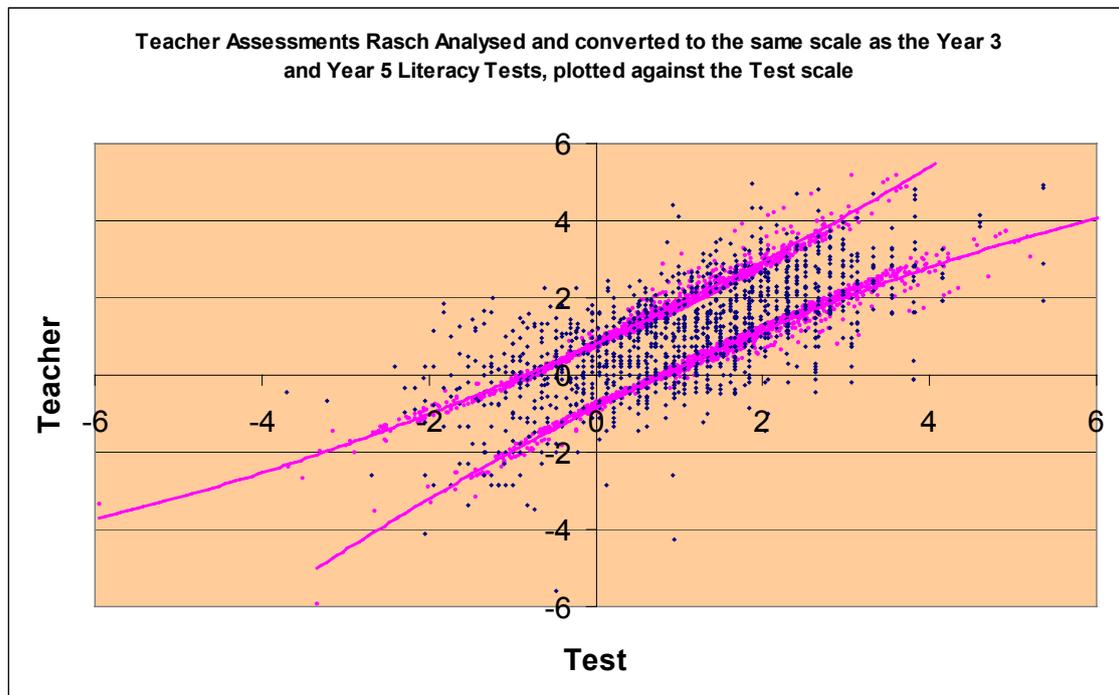
In summary just under 60% of the cases fall within the control lines. Because the data deliberately never identified which students belonged to which teacher it is difficult to estimate what proportion of teachers might be assessing on approximately the same dimension as the test. Additional analysis is planned to identify patterns at sites. An incomplete analysis suggests a significant site effect, implying that teachers at a number of

Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   8 of 16

sites assessed consistently with each other; a result which might have been expected at sites where effort had been put into coming to common understanding about the Profiles criteria and through moderation of judging practices. A small number of sites show a much more random pattern of test and teacher assessments. No conclusion can been drawn at this stage about the proportion of teachers that might be assessing approximately on the same basis as the tests, but the shape of the data profiles implies that many are close for at least some of their assessments.

Comparing Teacher Assessments for a number of Year levels with a model of Test data for the same Year levels.

An alternative process for examining teacher assessments is to explore the patterns of age 'signatures' within years and aggregated across year levels, to see how well these 'signatures' match those of tests. Tests have very clear age shapes within a year level. Grissom (2004) plotted test scores against age in months for a number of tests used in statewide testing in California. The SAT/9 Reading Test used in Grade 2 and with 455,638 cases, produces a pattern of test scores increasing for each month of age for each of the 12 months of the age normal peer group, that is the group that started the grade (and school) within the normal 12-month span. At the 13th month and for all subsequent months, the scores drop, indicating that those students above the normal age for the grade are performing at a lower level than those within the age normal peer range. The same general phenomenon is observed in the Tests used in this study. Only three Tests so far have had dates of birth attached to their data, Year 3 1997, Years 3 and 5 in 1998. All show approximately the same age 'signature', illustrated in Figure 3, Year 3 in 1997.
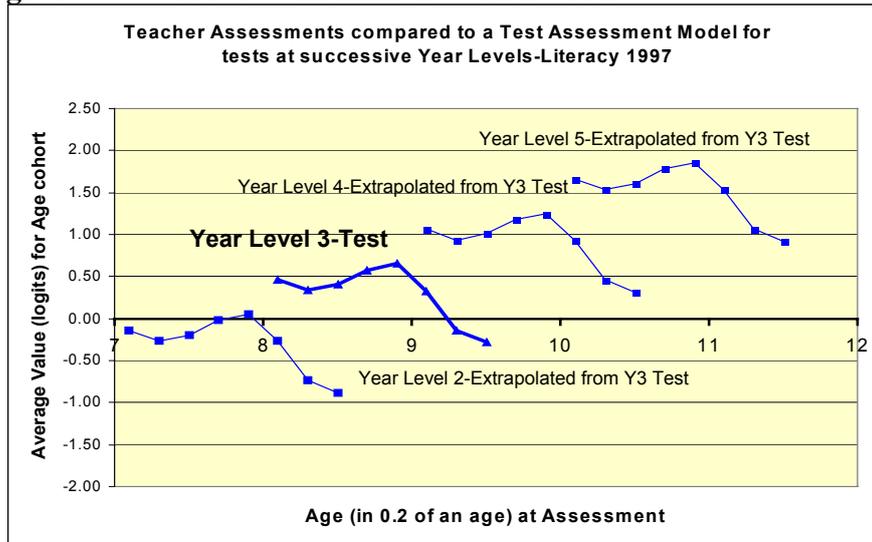
Figure 3   Test Age 'Signature', Year 3 Literacy, 1997
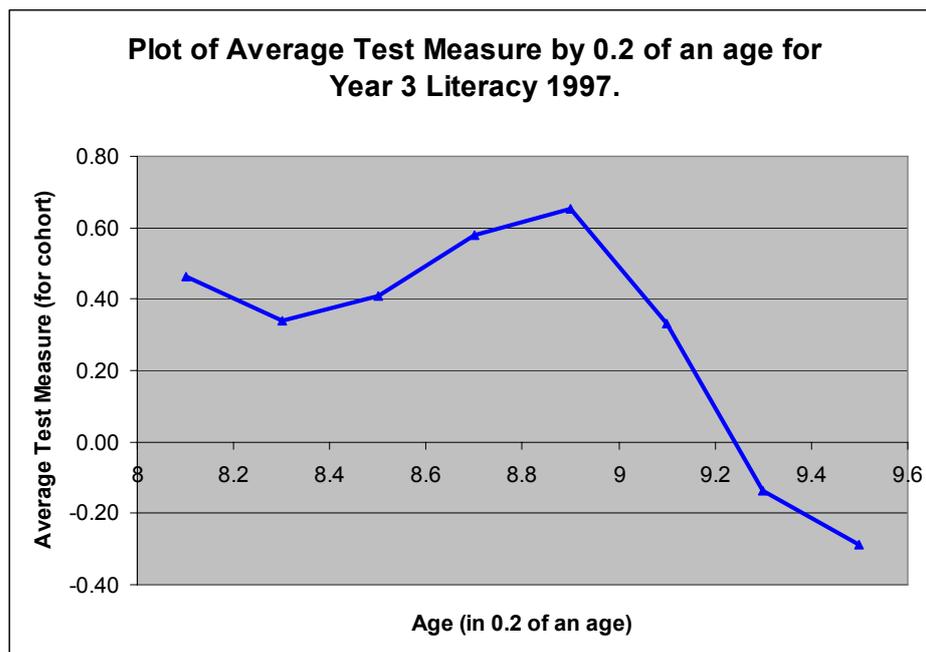
South Australian students show a small 'kick-up' in scores at the youngest age (for this year level, 8.1 years), but otherwise the pattern is very similar to the Californian data. While not ready for reporting yet, the pattern for Year 3 and 5 in 1998 show similar 'signature' patterns.

Using the general pattern for Year 3 a model of student growth (in cross-section) can be estimated by generating data for years 2, 4 and 5. Four replicates of 8988 cases for Year 3, modified to provide data for Years 2, 3,4 and 5 generate patterns as shown in Figures 4 and 5.

Figure 4 A Model of Test data for Years 2 to 5

Teacher Assessments compared to a Test Assessment Model for tests at successive Year Levels-Literacy 1997

The age and Literacy scores for each Year 3 case were systematically modified to create the data for years 2, 4 and 5. A Year 3 case becomes a Year 2 case by subtracting one year of age, a Year 4 case by adding one year, Year 5 by adding 2 years.

Real data by date of birth were not available for Year 5 at the point of analysis. The real age 'signature' for Year 5 (based on 1998 data where dates of birth have been matched) is slightly flatter for ages 10.3 to 10.9, than for Year 3. Given that Year 5 is the upper point of the model series, the reduced growth gradient does not strongly influence the pattern.

The measure, Literacy score in logits, is modified for each of Year 2, 4 and 5 by the estimated annual growth based on half the distance between the mean of the Year 3 and Year 5 total test populations, equated by the original test analysts. This annual value is 0.6 logits. The effect is to make the model show perfect linear growth by year level, an unlikely situation with real data.

Figure 5 A Model of Test Data for Years 2 to 5, across year levels, 1997

Test Assessment Model for tests at successive Year Levels accumulated for each age-Literacy 1997
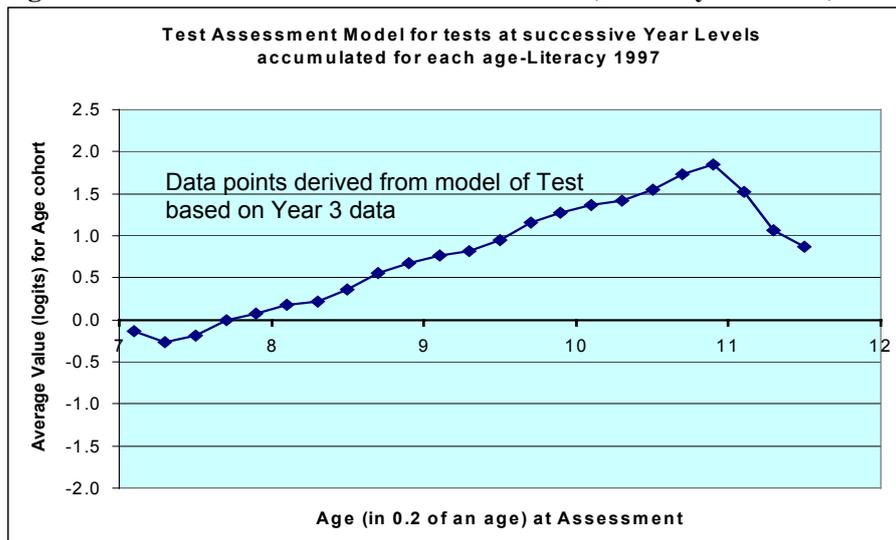
Figure 5 illustrates the effect of combining all the data from Year 2 to 5 by age. An approximately linear 'growth' curve is developed from each of the individual year level specific curves. The Test model is the result 1104 cases per age category by year level (Figure 4) on average and 1624 cases on average for each age category aggregated over Year 2 to 5 (Figure 5). These two models provide curves that can be compared to the Teacher Assessments.

Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   10 of 16

Figure 6 Teacher Assessments compared to a Test Assessment Model, 1997

Teacher Assessments compared to a Test Assessment Model for tests at successive Year Levels accumulated for each age-Literacy 1997
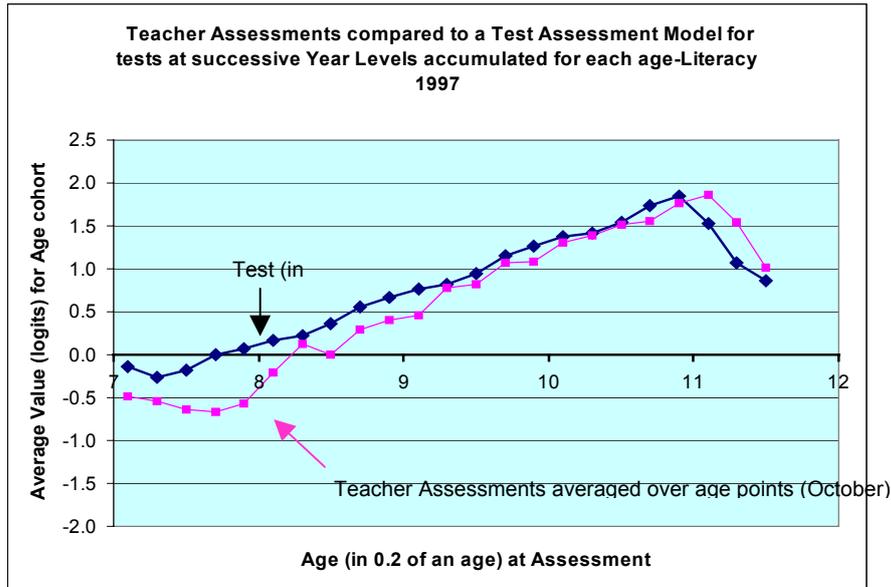
The Teacher data are based on an average of 115 cases per year level per age category and 194 cases for the four year levels combined per age category, as shown in Figure 6.

The data appear to match the model quite well but appear to be slightly out of phase. When a correction to the cohorts is made to adjust for the different assessment periods (test in August, Teacher in October), effectively to slide the then two month older Teacher assessed students back one cohort, to place them in the cohort group they would have occupied in August, the curves appear to coalesce, especially in the region of 9 to 11 years.

Figure 7 Teacher Assessments compared to a Test Assessment Model, corrected for collection months, 1997

Teacher Assessments compared to a Test Assessment Model for tests at successive Year Levels accumulated for each age-Literacy 1997
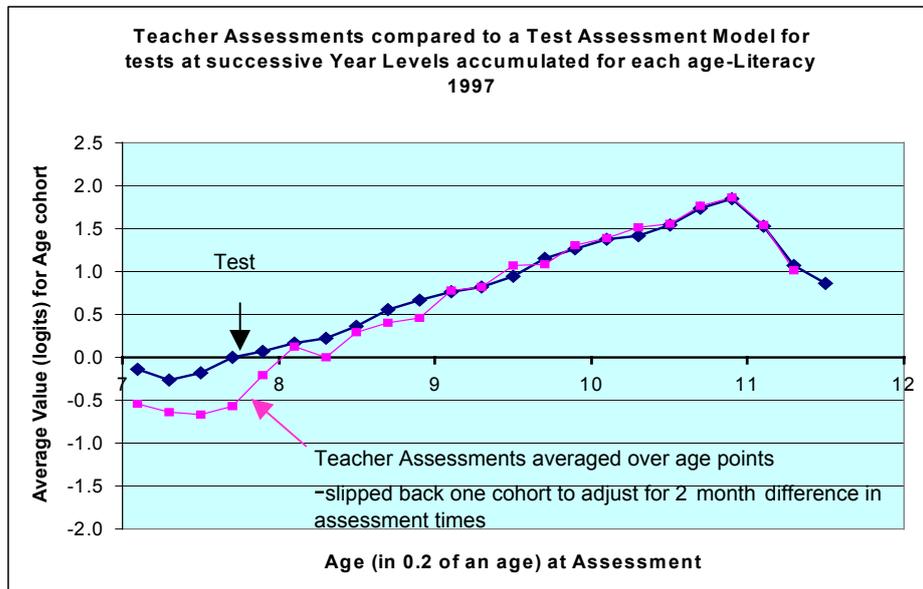
If the two curves are disassembled as in Figure 8. It is clear from the comparison of the two curves, now adjusted for the time-shift, that they are of very similar shapes.

Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   11 of 16

Figure 8 Teacher Assessments compared to a Test Assessment Model, corrected for collection times, by year level, 1997
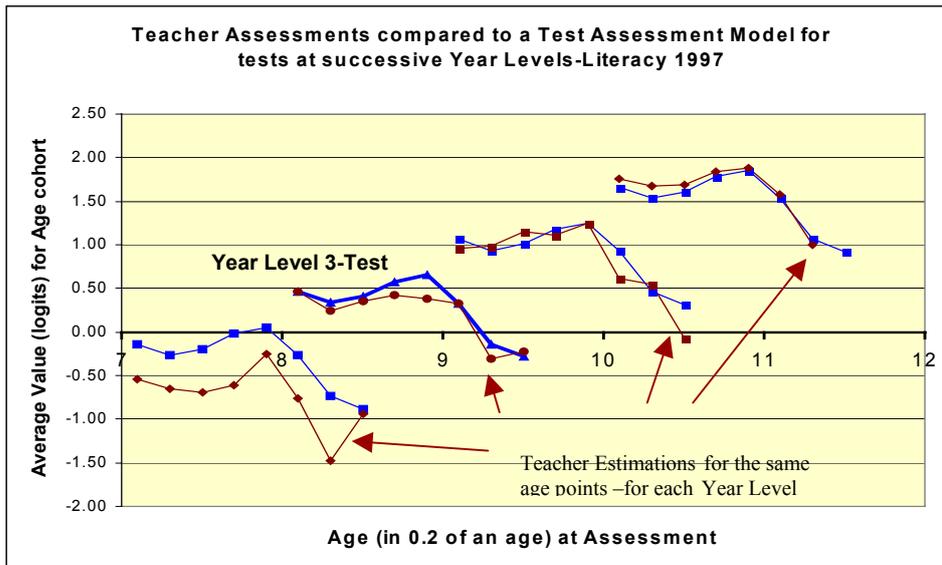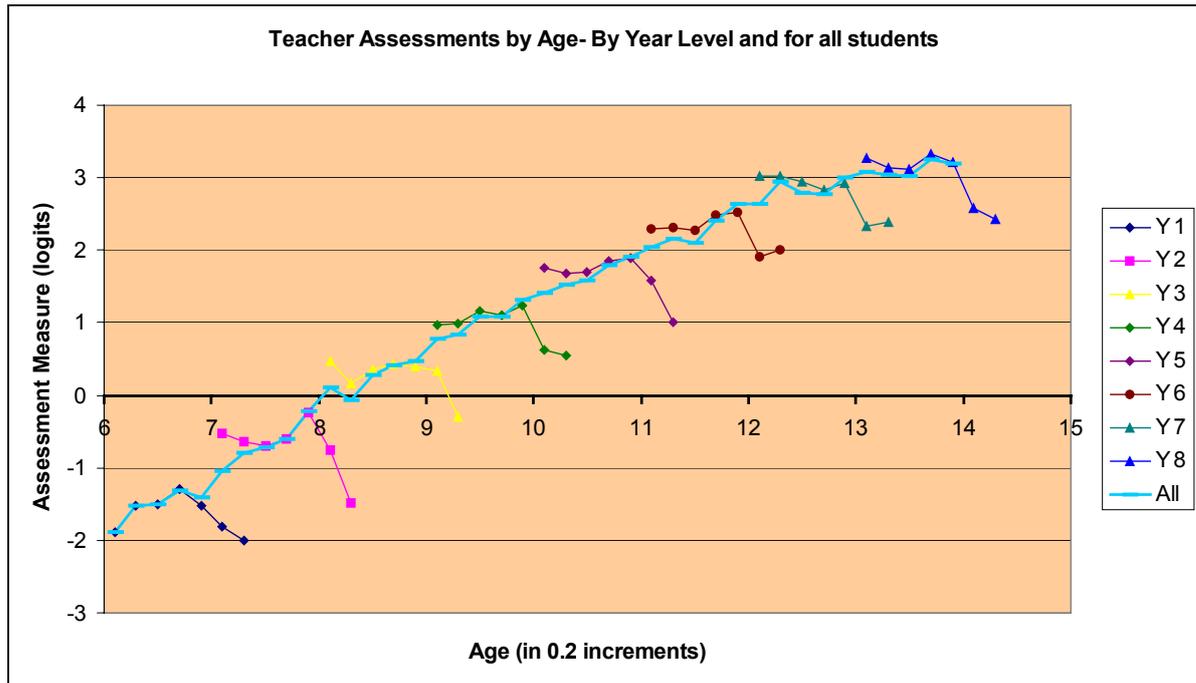

The Year 2 curves are further apart than are those for years 3, 4 and 5 of the Test model as the model assumes a constant growth factor (0.6 logits) from year level to year level. The independent Rasch analysis of the teacher assessments indicates a likely greater (steeper) growth from Year 2 to 3 than the constant growth model allows. The most remarkable feature is the apparent coincidence of the lines, and most importantly the phase shift required to make this happen. A match along the lines of the unadjusted data (Figure 6) would have been sufficient to suggest that enough teachers were assessing in ways consistent with the test for a general pattern of similarity to exist. That a shift of one age category (0.2 of an age) can make the match much closer, implies that enough teachers are assessing students on exactly the same basis as the tests and that the Teacher assessments are sensitive to about 0.2 of an age. Put another way, the aligned teachers are sensitive to changes in student learning at about 0.1 of a logit, the improvement in the 'average' student in a period of about 8 weeks.

Could this be some serendipitous alignment? The data for 1998 for Mathematics have not yet been analysed fully but nearly exactly the same patterns exist when the teacher assessments are matched by age to a model of test data. There is one exception. Because the collections of data for both tests and teachers were conducted during the same period no adjustment to the curves is required, that is they seem completely in phase.

Figure 9 shows the full spectrum of the average of teacher assessments from Year 1 to Year 8. The general age 'signature' is retained at all year levels. The curve of the aggregated data is less uniformly linear than the artificial test model, based on two key points only; Year 3 and Year 5. Higher year levels show a diminished within-year age effect, that is the within-year gradient of change by age, flattens requiring the aggregate curve to flatten. This reflects the change at higher levels reported by Grissom (2004) where the within normal peer age effect flattens by Grade 10. The transition years from primary to secondary levels show lower growth rates. Further investigation will establish whether tests at these year levels show a similar pattern.

Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   12 of 16

Figure 9 Teacher Assessments by Age, by Year Level, 1997.

Discussion

It is rare to come across data sets that attempt to assess a wide range of student learning at different year levels on a common basis. While untested for validity or reliability at the time of its use, the progress indicator provided an additional dimension for consideration in student assessments. Original summaries of the data in 1998 and 1999 provided insights from a cross-sectional analysis, of what general growth patterns for students might look like. These patterns were not fully explored at the time, and with the replacement of the Profiles for Australian Schools by the SA Curriculum and Assessment Framework, this work was put to one side.

The ongoing analysis described in this paper explores what indicators for further development of classroom assessment might have come from a deeper exploration of the subtle patterns in the data.

The survey procedures ensured that no teacher details were ever collected so the link to general patterns of traditional teacher characteristics (gender, age, experience, attitude, pedagogic approach etc) cannot be explored. Which students were taught by which teacher was not collected, so an analysis by teacher is impossible. The analysis of Teacher assessments using the Rasch model, while stretching some of the requirements of the model, has provided a mechanism to convert teacher ratings of students into measures. The measures have been aligned, approximately, to the Test measure based on 1200 of the 8000 students in the teacher survey, who also sat the Year 3 and Year 5 tests in 1997.

The 1200 cases common to the Test and Teacher assessments have a correlation coefficient of 0.66 indicating a reasonable pattern of matching of teacher assessments to test scores. Bringing teacher assessments to the same scale as the tests show about 57% of assessments can be regarded as identical. Not yet completed analyses by worksites show patterns of high consistency of test and teacher assessments at some sites and random patterns at other sites. This implies that some teachers were relatively consistently assessing students on approximately the same scale as the test, others were assessing on some other basis. Enough teacher assessments matched the test assessment to warrant explorations of other patterns in the data.

Comparison of Teacher assessments by age, in 0.2 of an age, show a pattern of age 'signature' similar to that of a test. By building a data set from the Year 3 test data to simulate a data set

Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   13 of 16

for years 2 to 5, it has been possible to show that the average of teacher assessments at 0.2 of an age provide a very similar growth pattern to that displayed by the simulated test data. The resultant comparison shows two growth curves slightly out of phase. When the Teacher assessments are corrected for the different ages at test assessment (August) and teacher assessment (October), the two curves coincide for a range of ages from 9 to 11.

When adjacent age cohorts are compared for growth in either the model of test data, or in the more variable Teacher assessment data, there is no statistically significant difference and effect sizes of growth, even for a full 12 months, are very small. Effect sizes are 0.15 to 0.20 for growth by year level or by normal age (the integer value of a students age- age as it is commonly understood), and of the order of 0.04 to 0.06 for a 0.2 age increment. This is 'very small' (as described by Izard, 2004 based on Cohen, 1969) as a magnitude of effect size. Nevertheless the general pattern of 'growth' can be seen in the teacher data aggregated for all year levels at a given age, and in the pattern for each year level by 0.2 of an age.

The coalescence of the Teacher and Test curves when an adjustment for test date is made, implies a reasonably sensitive measure by that set of teachers approximating the test scale. Their number is unknown but we assume it might be of the order of 20 to 40 percent of the cases, since a reasonable number of cases must apply to produce the effect. That this is not mere aberration is confirmed in the not yet fully analysed 1998 data, where common assessment dates appear to produce overlapping, in-phase 'growth' curves without adjustment. The replication of the phenomenon within the 1997 data for most year levels, and the quite independent replication of the phenomenon in the 1998 data at most year levels, suggests that a reasonable number of teachers were assessing on the same latent dimension as the test, and with a very high degree of sensitivity. That sensitivity is estimated to be about 0.1 of a logit, equivalent to noticing a measurable change in a student's learning over an eight-week period. Could we expect Teachers to be even more sensitive observers of change?

School administrations and their psychometricians might have done a disservice to teachers by not developing adequately refined assessment tools and insights for teachers, to encourage them to record assessments at the level of sensitivity that appears possible. South Australia chose not to explore the possibilities of scale-based recording and Victoria has used too coarse a scale. Victorian teachers have continued to use a three-zone assessment scale within major criteria, requiring average growth of about 0.3 logits, of the order of six to nine months' growth. This lack of refinement is one of the reasons parents have been dissatisfied with criterion based assessments, and have supported the move back to alphabetical grades. Victoria has, significantly, in comparison to other States, retained elements of its vision for student assessment in its proposed grading process, choosing to use intervals of 6 months above and below the standard for a year, consistent with the 0.3 logit estimation above, and making this the basis of grades (see Victoria- Sample Report Card, 2005).

Analysis of the level of sensitivity at which teachers can discriminate changes in students' learning might lead to more useful recording processes for documenting teachers' observations of learning, based on their judgements, fine-tuned by deeper analysis of the difference between test results and teachers' assessments. New research investigating the ability of teachers to predict test scores for students in the cohorts where testing occurs (Yrs 3,5,7) would both inform the research and school communities of the degree to which current teachers can match test assessments, and automatically set up the process for teachers to refine either their judgements through supportive feedback, or to engage in dialogue with test designers about any obvious mis-measured cases in a mandated test.

Enough teachers assessed close to the same basis as the test, to ensure that the patterns of learning development by age, as observed by teachers, matched the patterns of learning displayed by a model based on tests. Distilling from the most effective assessors how they do it might offer a better development path for classroom assessment than mass testing.

The result described occurred even though descriptions of the criteria (Profiles, CSF) were ambiguous and open to varied interpretations, making a consistent understanding of any given level elusive. The work of Forster and Masters (Forster & Masters, 1996, Masters & Forster, 1996) and the recently published NAEP 2005 Math Item Maps (NAEP, 2005) among other

Paper Presented at the AARE Annual Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   14 of 16

'progress maps', that highlight the average sequences of skill development and the relative difficulties of any skill, are critical tools to assist teachers with their observations. Secondly, further research on the techniques to refine teacher assessments for them confidently to place individual students on developmental scales at any time, would enable 'on-balance' teacher assessments to be used as a basis of growth focused education. Development from time x to time x+1, would then be reinforcing, as part of formative assessment, for students, teachers and their parents and other interested parties. Continuity of assessment schemes across Year Levels would be feasible also, facilitating the growth perspective. A further key benefit of effective scale placement by teachers is the ability to interpret a position at any time as 'does these things' and 'does not yet do these things consistently'.

A focus on learning 'latent dimensions' for key skills and knowledge as a basis for developing curriculum descriptions, would begin to provide the additional tools needed by teachers to estimate better any student's developmental position at any time. These tools would also include standardised probing strategies, where individual student responses could be indicators of likely scale position. Tests, especially computer adaptive tests will help but enhancing the teachers' ability to observe and interpret students behaviour and responses in terms of where their current skill level might lie will be particularly useful. If data from all tools were convertible to common scales, as, for instance, in the Lexiles (Stenner and Stone, 2004) and Quantiles attempts, over time a new language of supportive description of educational progress might evolve, replacing C and Ds and Es with 'monthly or weekly development' expressed in some form of common scale units, and thus help teachers keep focused on the 'zone of proximal development' for each student.

The hope of our analyses is that it provides evidence to support the development of teachers as trusted assessors and the integrators of data about students from multiple sources, without requiring expensive 'buy-in' to any given assessment or text book schemes. Teachers are our most critical resource in the educational process and must not be relegated to the role of mere technician. Improving teachers' abilities to make 'on-balance' assessments on scales of development with high reliability and consistency will be much more cost effective than providing more and more high stakes assessment technology and pressure at the classroom level. Confidence in teachers' abilities to do this will improve the more teachers can be shown to be able to estimate students' test scores with high accuracy. This accuracy will improve partly on the basis of good regular feedback on estimation accuracy. Ideally an interactive system of testing and teacher assessment will be developed that keeps everyone assured of quality assessments, helps teachers and psychometricians recalibrate their judgements and focuses on individual student learning growth.

References

Associated Press (2005), *Some States to Get Wider Latitude in Measuring Students' Gains*, New York Times, November 19, 2005

Bates, C. & Nettelbeck, T (2001). *Primary School Teacher's Judgements of Reading Achievement*. Educational Psychology, 21, 2, 177-187.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8–21.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–147.

Bond, T. & Bond, M., (2003*), Measure for Measure: Curriculum Requirements and Children's Achievement in Music Education*, Paper presented to the AARE Annual Conference, Auckland, Nov 2003.

Brookhart, S.M. (2003). *Development measurement theory for classroom assessment purposes and uses*. Educational Measurement: Issues and Practice 22(4), 5-12.

Brookhart, S.M. (2004). *Grading*. Pearson Education Inc.. New Jersy.

Coladarci, T. (1986). *Accuracy of Teacher Judgments of Student Responses to Standardized Test Items*, Journal of Educational Psychology, 78, 2, 141-146.

Consistency of Teacher Judgement CD-ROM, (2000), South Australian Department of Education, Training and Employment with Vic. Dept of Ed., Employment and Training and the Queensland School Curriculum Council, Copyright Commonwealth of Australia.

Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   15 of 16

Cronin, J., Kingsbury, G.G., McCall, M.S., Bowe, B. (2005). *The Impact of the No Child Left Behind Act on Student Achievement and Growth: 2005 Edition*. Northwest Evaluation Association

Curriculum Corporation (1994) *The Statements and Profiles for Australian Schools*. Melbourne. Curriculum Corporation. (Set of 8 documents published for Australian Education Council).

Forster, M., & Masters, G (1996). *Portfolios Assessment Resource Kit (ARK Portfolios)*, The Australian Council for Educational Research Ltd..

Fuller, M. (2000). *Teacher Judgment as Formative and Predictive Assessment of Student Performance on Ohio's Forth and Sixth Grade Proficiency Tests*, paper presented to the AERA Annual Meeting, April 2000, ED 441 015.

Griffin, P. (2004). *The comfort of competence and the uncertainty of assessment*. Paper presented at the Hong Kong School Principal's Conference, Hong Kong Institute of Education, March.

Grissom, J.B. (2004). *Age and Achievement*. Education Policy Analysis Archives, 12 (49). Retrieved from http://epaa.asu.edu/epaa/v12n49/.

Guskey, T. R. (1996). *Reporting on student learning: lessons from the past, prescriptions for the future*. ASCD Yearbook, Chapt 3.: http://www.ascd.org/readingroom/books/guskey96book.html#chapter3

Hauser, C. (2003). *So, what d'ya expect? Pursuing individual student growth targets to improve accountability systems*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Hungi, N.(2003). *Measuring School Effects Across Grades*, Flinders University Institute of International Education Research Collection, Number 6.

Izard, J. F. (2004, March). *Best practice in assessment for learning*. Paper presented at the Third Conference of the Association of Commonwealth Examinations and Accreditation Bodies on *Redefining the Roles of Educational Assessment*, South Pacific Board for Educational Assessment, Nadi, Fiji.

Keeves, J. & Majoribanks, K., (eds) (1999). *Australian Education: Review of Research 1965-1998*, ACER

Kingsbury, G G., Olson, A., Cronin, J., Hauser, C., Houser, R., (2004), *The State of State Standards: Research Investigating Proficiency Levels in Fourteen States*. Northwest Evaluation Association, http://www.young-roehr.com/nwea/

Kingsbury, G. G. (2000). *The metric is the measure: A procedure for measuring success for all students*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Marzano, R. J (2000). *Transforming Classroom Grading*, Association for Supervision and Curriculum Development.

Marzano, R. J (2003). *What works in schools: Translating research into action*, Association for Supervision and Curriculum Development.

Marzano, R. J. (1998). *Models of standards implementation: Implications for the classroom*. .http://www.mcrel.org/products/standards/models.asp

Masters, G & Forster, M.(undated)   *The Assessments We Need*, ACER Website, www.acer.edu.au/research/ documents/Theassessmentsweneed.pdf

Masters, G. N. (2004). *Continuity and Growth: Key Considerations in Educational Improvement and Accountability*. Address to the joint ACE and ACEL National Conference, Perth, October.

Masters, G., & Forster, M (1996). *Developmental Assessment: Assessment Resource Kit (ARK Developmental Assessment)*, The Australian Council for Educational Research Ltd..

Meisels, S.J., Bickel, D., Nicholson, J., Xue, Y., & Atkins-Burnett, (2001), *Trusting Teachers' Judgments: A Validity Study of a Curriculum-Embedded Performance Assessment in Kindergarten to Grade 3*. American Education Research Journal, 38, 1, pp 73-95.

National Research Council (2003). *Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment. Workshop report.* Committee on Assessment in Support of Instruction and Learning. Board on Testing and Assessment, Committee on Science Education K-12, Mathematical Sciences Education Board. Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

NAEP (2005). *Map of Selected Item Descriptions on the NAEP Mathematics Scale — Grade 4,* http://nces.ed.gov/nationsreportcard/itemmaps/index.asp?subj=Mathematics

Office of Review, *Benchmarks 96* and *Benchmarks 97* series, (1997), Department of Education, Victoria, Australia

Profile Data for English, Science, Mathematics, Health, Physical Education Participation and Achievement Series: Nos 2-5, 1998, 1999, Department of Education, Training and Employment, South Australia.

Rothman, S. (1998). *Factors Influencing Assigned Student Achievement Levels*. Paper presented at the AARE Conference, Adelaide, November.

Rothman, S. (1999). *Factors Influencing Assigned Student Achievement Levels II*: Mathematics*, The Arts, and Health and Physical Education*. Paper presented at the joint AARE & NZARE Conference, Melbourne, November.

Rowe, K.J.& Hill, P.W. (1994). *Assessing, recording and reporting students' educational progress: The case for 'Subject Profiles'*. Assessment in Education, 3,pp 309-351.

Paper Presented at the AARE Annual
Conference, Parramatta 2005

Bond, T & Caust, M. (2005). Silk purses from sows' ears? Making measures from teacher judgements,   AARE Sydney, BON05361   16 of 16

Stenner, A. J. & Stone, M. H. (2004), *Does the Reader Comprehend the Text Because the Reader Is Able or Because the Text Is Easy?* Paper presented at International Reading Association Reno-Tahoe, Nevada May 4, 2004 From Metametrics Website www.lexiles.com

Stiggins, R. J. (2001). *The unfulfilled promise of classroom assessment*. Educational Measurement, Issues, and Practice, 20, 3, 5-15.

Victoria. Sample Report Card, 2005. http://www.sofweb.vic.edu.au/studentreports/pdfs/primary_example.pdf

Wilson, M, (2004) *Assessment, Accountability and the Classroom: A Community of Judgment*. in Wilson, M. (Ed.).(2004). *Towards Coherence Between Classroom Assessment and Accountability*, 103rd Yearbook of the National Society for the Study of Education, University of Chicago Press, Chicago.