

ResearchOnline@JCU

This file is part of the following reference:

Caust, Martin Kennings (2010) *Measuring student progress in school: a role for teacher judgement*. PhD thesis, James Cook University.

Access to this file is available from:

<http://eprints.jcu.edu.au/18998>

The author has certified to JCU that they have made a reasonable effort to gain permission and acknowledge the owner of any third party copyright material included in this document. If you believe that this is not the case, please contact ResearchOnline@jcu.edu.au and quote <http://eprints.jcu.edu.au/18998>

**Measuring student progress in school:
A role for teacher judgement**

Thesis submitted by

Martin Kennings CAUST BSc Adelaide, BSc (Hons) Qld

September 2010

for the degree of Doctor of Philosophy

in the School of Education

James Cook University

Statement of access

I, the undersigned, the author of this thesis, understand that James Cook University will make it available for use within the University Library and, via the Australian Digital Theses Network for use elsewhere.

I understand that as an unpublished work, a thesis has significant protection under the Copyright Act and I do not wish to place any restriction on access to this thesis.

14/12/2010

Signature

Date

Contribution of others

Financial Support

Financial support was gratefully received from the James Cook University School of Education, through Professor Trevor Bond, for the initial analysis of the data.

A travel grant was provided by the School of Education to attend the 2005 Pacific Rim Objective Measurement Symposium in Kuala Lumpur to present a paper.

Supervision

Professor Trevor Bond was principal supervisor.

Dr Helen Boon was second supervisor.

Editorial Support

Professor Bond and Dr Boon provided editorial advice as part of their supervision duties.

Access to data

The Department of Education and Children's Services, South Australia agreed to the release of data for a re-analysis (see Appendix 1).

Statistical Support

Professor Bond provided detailed advice on the use of the Rasch model.

Statement on sources

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education.

Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Signature

14/12/2010

Date

Acknowledgements

I am deeply grateful for the encouragement and support provided by my partner Margie Darcy over what turned out to be a longer period than expected. Without her positive approach in the face of a diminished social, family and work life, this document would not exist. She very generously accepted the limitations imposed by a student life.

I am appreciative of the support from Alan Green, of the Department of Education and Children's Services, in clearing the way for the release of data. The data are unique in South Australian education, and while parts were widely reported in the years of their creation, the opportunity to explore them for other messages has been rewarding and I believe tells us that the professional role of teachers as observers of learning is under-utilised. The assistance of Gary O'Neill in the creation of the data files is greatly appreciated.

The data would not exist in the form analysed without the conceptual and software development of Ian Probyn who was able to create an intuitively simple data collection process. He has also provided a sympathetic ear to the ongoing summary of what *his* data seem to say. The writing project would not have commenced without personal references from John Ainley and Richard Jenkin, nor without an initial encouragement possibly now long forgotten, from Professor John Keeves.

Articulating the concepts in the thesis was made possible through the strong support and ongoing reassurance of Professor Trevor Bond. He graciously agreed to supervise a complete stranger who tracked him down through the wonders of the internet. A more encouraging supervisor would be difficult to find. His openness to a less conventional approach and his non-pedantic and sensitive supervision are deeply appreciated.

Dr Helen Boon came late to the supervisory challenge of helping a slow and ponderous writer complete a manuscript. I am grateful she was prepared to wade through the draft chapters to improve their logic and readability. Neither supervisor is responsible if any errors of logic or fact remain.

Finally to our children, all four of them and their partners, my four sisters and my friends in Adelaide, Canberra and Sydney who suffered with or because of me, I acknowledge your support and regular inquiries about progress. To Sue Pender, a regular encourager and mentor, a particular thankyou.

I dedicate the thesis to Tess Caust, who showed it is never too late to study and that a passion for something can make a difference.

Abstract

The documentation of learning is a weakness of all schools and systems, leading to complaints about the lack of information and a press for teacher accountability. Current solutions to increase information about learning and improve accountability promote standardised (and national) testing of student cohorts and/or better use of often-archaic classroom assessment results. System-wide testing, while not without value for some purposes, is very limited in its contribution to improving classroom practice. In particular testing is a process detached from the needs of classroom teachers and given the time for results to be returned, unhelpful in timely decision making.

Assessment of students by teacher judgement is a general feature of classroom teaching but its quality is often unknown. This thesis addresses the history and application of teacher judgement assessment and then analyses teacher and test assessments of the same populations of students (from South Australia in 1997 and 1998). The analyses establish the comparability of the assessment processes, and thus one basis for inferring the quality of teacher judgement. The purpose is to test the feasibility of using teacher judgement assessments, calibrated to scales of learning, as the prime data to record, manage and report learning and monitor its change over time.

In curricula structured in levels, as apply in some Australian school systems, one possibility for recording assessments is in the form of the level judged to be most recently achieved. Over an extended time frame a general trajectory of learning for each student can be documented. If the progress made as a student learns new skills, knowledge and understandings could be assessed and recorded by a teacher in finer detail than a level, a basis might exist for documenting learning with utility for teachers, students and all other parties interested in being kept informed. These two broad ideas, the teacher's concept of learning in a specific strand of the curriculum and the mandated test as one method to describing that learning, are brought together to appraise the feasibility of creating methods of assessing and recording learning, built upon the constructs rather than any particular test or assessment process.

The data analysed are unique. They are limited to two calendar years (1997 and 1998) for two learning areas and are useful in estimating the potential for teacher and test assessments to track the learning development of students over time in the same fashion. Within the limitations of the data the potential of teachers to record the learning development of students directly, using broad scales to locate their current learning status is confirmed. Very strong similarities are found in the general characteristics of the data once the teacher scale is transformed to the scale of the test. Both assessment processes show increments in mean leaning for age cohorts grouped in 0.1 of year of age and smooth growth trajectories with age and Year level. Both processes show marked gender differences for English language, trivial gender differences for mathematics. Both processes show

within Year level patterns by age and gender that are consistent with test data analyses found elsewhere.

When case studies for individual schools are examined, it is clear that at some sites teachers assess with high correlation to the test scores, indicating the potential for easily recalibrating some teachers to increase the match of the assessments from the two processes. It appears potentially feasible to design classroom and school assessment systems on the basis of teacher judgement assessment data as the prime data source. Test data can be integrated readily and usefully into the scheme. The issues that need further consideration are outlined along with the general implications for support to teachers, training and re-training and some broader data management issues for classrooms, schools and systems. Subject to the resolution of a number of design issues, schools and school systems might then optimise the skills of teachers as both managers and documenters of learning. This would allow for the professional skills of teachers to be acknowledged and capitalised upon. Rather than the assessment skills of teachers being directly derided, or derided by implication as a consequence of externally imposed testing procedures, testing arrangements might be reconfigured to support and confirm the quality of teacher judgement assessments.

Table of Contents

Statement of access	ii
Contribution of others	iii
Statement on sources	iv
Acknowledgements	v
Abstract	vi
Table of Contents	viii
List of Tables	xii
List of Figures	xiii
CHAPTER 1: MAPPING THE TOPICS OF INTEREST	1
Overview of this chapter	1
Overview of subsequent chapters	2
Elaborating the main character: Teacher Judgement Assessment	3
Assessment as a support to learning	7
Propositions considered	9
Key questions considered	10
Learning: an operational definition for this thesis	11
Progression	14
Personalised learning	16
Understanding learning with student test data	17
Strengths and limitations of the study	17
Summary	18
CHAPTER 2: EARLY APPROACHES TO QUANTIFICATION OF LEARNING AND SCALE DEVELOPMENT	20
Timeline of the key examples considered	21
1845 Massachusetts: the first system wide examination process in the US	22
1850-1862 Fisher Scale Book and numerical approach	23
1892-1908 Rice's educational surveys and Stone's enhancements	25
1909-1911 Courtis and the influence of Stone	27
1910 Thorndike and the handwriting scale	32
1912 Thorndike's concept of scaling	34
1912 Hillegas: judging the quality of prose	35
1913 Criticisms of scaled approaches from researchers of the period	38
1914-1916 Thorndike's scaling for reading	39
1916 Trabue's completion test	40

Item Difficulty and the key link to educational measurement	41
1916 A ‘level’ approach for composition	42
1950s - A new way forward.	43
Summary	44
CHAPTER 3: LEVELLED CURRICULA, LEARNING PROGRESS AND SKILLS TESTS	47
The development of ‘Profiles’ and ‘Levels’ for Australia	47
Implementation in South Australia	50
Confirmation of the value of profiles – application in studies and student assessment	55
Criticisms of a level approach	57
A Parallel Universe - the Testing Approach	58
Summary	59
CHAPTER 4: TEACHER JUDGEMENT ASSESSMENT– ISSUES, METHODS, AND CASE STUDIES	61
Issues in comparing teacher judgement and test assessments	62
Methods comparison	65
Clarifying how teachers make judgement assessments.	71
Studies/examples of the use of teacher judgement in research and classroom practice	74
Do accurate teacher assessments influence learning?	108
A summative overview of teacher judgement -Harlen	110
Summary	111
CHAPTER 5: THE TRAJECTORIES OF LEARNING, GROWTH AND GROWTH INDICATORS	114
Establishing trajectories of learning growth with age and Year level	115
Learning growth in cohorts - examples of growth trajectories for the test score means of groups of students	123
Patterns by age within Year level	137
Case studies where further analysis of test data might provide scaled indicators of student development	141
Comments on individual learning trajectories	145
Summary	147
CHAPTER 6: SOUTH AUSTRALIAN TEST DATA FOR 1997 AND 1998	149
Literacy and Numeracy Tests	149
Rasch model analysis of the Literacy and Numeracy tests	150

The trajectory of Literacy test scores	153
The trajectory of Numeracy test scores	165
Overview of the Literacy and Numeracy test models	173
Summary	175
CHAPTER 7: SOUTH AUSTRALIAN TEACHER JUDGEMENT ASSESSMENTS: 1997 AND 1998	177
The data collection revisited	177
The English Learning Area	179
The Mathematics Learning Area	186
Common findings across two data collection periods and learning areas	194
Acceptability of teacher judgement assessment to teachers	195
Concluding comments	196
CHAPTER 8 TEACHER AND TEST ASSESSMENT COMPARED	198
Equating Teacher and Test scales	198
Comparing Teacher and Test Assessments for Common Students with Teacher Assessments Re-scaled.	212
Extending the comparison of Teacher and Test assessments beyond Years 3 to 5	224
Is the variability in assessment alignment a within-teacher or between-teacher effect?	236
Summary	237
CHAPTER 9 WEAVING THE THREADS TOGETHER	241
Appraising the principal character	241
The main findings from the data analysis	241
The propositions: findings	244
Responses to questions posed in Chapter 1	245
Design elements for a teacher judgement assessment scheme	250
Addressing the remaining questions from Chapter 1	256
In conclusion - the fate of the principal character	266
REFERENCES	270
Appendix 1 Letter of approval for data access	294
Appendix 2 Adult literacy trends with age	295
Appendix 3 Adequacy of the Key Stage Test Assessments	297
Appendix 4 Scale changes CSF to VELS in Victoria	299
Appendix 5 NAPLAN data and model.	301
Appendix 6 North West Evaluation Association -Data confirming learning trajectories	312

Appendix 7 Mathematics Assessment for Learning and Teaching (MaLT) in England	314
Appendix 8 Curriculum, Evaluation and Management (CEM) Centre-Consistency in the learning difficulty Scale for numerals as an example of potential tools to support teachers.	317
Appendix 9 Chicago-Strategic Teaching and Evaluation of Progress (STEP)	325
Appendix 10 Individual learning trajectories.	330
Appendix 11- Summary of equating approaches and issues	340
Appendix 12 Summary of Rasch analysis statistics for teacher judgement assessment data	342
Appendix 13. Estimates of the proportions of teachers at various levels of correlation and match to the test scales.	349

List of Tables

Table 2.1 Timeline of historical developments in assessment considered	21
Table 4.1 Table of Kappa values	67
Table 4.2 Summary of Matches of Teacher and Test Assessments -Worcestershire LEA; data for 1997 to 2001 combined, with level as unit of reporting.	86
Table 4.3 Matches of Grand Average Teacher and Test Assessments-Worcestershire LEA, 1997 to 2001, with 1/3 level as unit of reporting	87
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.	129
Table 6.1 Students in the Basic Skills Test Program (BSTP) included in data analysis	150
Table 6.2 Summary of Winsteps Fit and Measurement Statistics, Literacy 1997	151
Table 6.3 Summary of Winsteps Fit and Measurement Statistics, Numeracy 1998	151
Table 6.4 Literacy – Mean scores by Year level and Testing Year	155
Table 6.5 Literacy-Comparison of original records with subsets assigned dates of birth	160
Table 6.6 Comparison of 2001, 2002 and 2004 Literacy score statistics - full cohorts	161
Table 6.7 Literacy Model-main statistical characteristics	162
Table 6.8 Numeracy – Mean scores by Year level and Testing Year	166
Table 6.9 Numeracy-comparison of original records with subsets assigned dates of birth	169
Table 6.10 Numeracy Model-main statistical characteristics	169
Table 7.1 English Learning Area by Year level–1997: General Statistics	181
Table 7.2 Mathematics Learning data by Year level –1998: General Statistics	190
Table 7.3 Ratings by teachers of their confidence in the process and in their specific assessments.	196
Table 8.1 Correlations of English teacher assessments with Literacy test assessments – 1997	200
Table 8.2 Correlations of Mathematics teacher assessments with Numeracy test assessments-1998	200
Table 8.3 General Statistical Characteristics of common cases of Teacher assessments and Test assessments, 1997 and 1998	201
Table 8.4 1997 Comparison of Teacher and Test assessments of common students	214
Table 8.5 1998 Comparison of Teacher and Test assessments of common students	215
Table 8.6 1997 “Deming” Regression and Kappa values for Teacher and Test assessments of common students at selected sites	219
Table 8.7 1998 “Deming” Regression and Kappa values for Teacher and Test assessments of common students at selected sites	222
Table 8.8 Estimates of the percentage of teachers in categories of correlation with the tests cross-tabulated with the rate of match to the test-1997 and 1998 data combined	236

List of Figures

Figure 2.1 Points Scores for Correct Steps-Fundamentals v Reasoning	29
Figure 2.2 Trabue’s Completion Test from Engelhard (1982)	42
Figure 4.1 Time Series of Teacher Assessments (TA) compared with Test Assessments. Percentage achieving at or above Level 4 for 11 year olds (Key Stage 2)-England	81
Figure 4.2 Teacher Assessments (TA) compared with Test Assessments. (2008), by Local Authority (LA). Percentages achieving at or above Level 4 and Level 5 for 11 year olds (Key Stage 2) England	83
Figure 4.3 Comparison of Times series of Year 3 Teacher and Tests Data (values estimated from original graphs in Victorian Auditor-General, 2009)	94
Figure 4.4 Comparison of Times series of Year 3, 5 and 7 Teacher and Test Data –Reading (values estimated from original graphs in Victorian Auditor-General, 2009)	95
Figure 4.5 Comparison of Times series of Years P-10 Teacher Assessment Data –Mathematics (Number 1-6/Chance and Data 7-10), by gender	97
Figure 4.6 Reading: All students-Mean Teacher Judgement Assessments 1999-2005 by Year level.	99
Figure 4.7 Reading: All students-Plot of regression parameters for each year 1999 to 2005.	100
Figure 5.1 Physical Growth Curve of American Males-median curve.	118
Figure 5.2 Model of NAPLAN Reading 2008 with indication of spread of data	126
Figure 5.3 Comparison of effect sizes at each Year level-NAPLAN, US	130
Figure 5.4 NWEA Reading Norms data (2002) with fitted curves	132
Figure 5.5 Effect size estimates for NWEA, NAPLAN and general US norms for Reading	133
Figure 5.6 Model of Mathematics Development - Mathematics Assessment for Learning and Teaching, (MaLT)	134
Figure 5.7 Effect sizes for Mathematics Assessment for Learning and Teaching compared with pooled US tests	135
Figure 5.8 SAT 9 Reading Scores Grade 2 (2002)- from Grissom (2004, p. 6)	138
Figure 5.9 Reading Test scores Early Childhood Longitudinal Study (ECLS) by age at testing	139
Figure 5.10 Numbers in Estimated Order of Difficulty to Say Aloud-all numbers to 20, samples from thereon (Difficulties relative to ‘1’)	143
Figure 5.11 Overview of STEP Letter Identification and Letter Sound Item Maps (from Figure 5 Kerbow & Bryk, 2005)	144
Figure 6.1 Literacy mean scores –Cross-sectional view with model trajectory	158
Figure 6.2 Literacy mean scores –Longitudinal view with model trajectory	159
Figure 6.3 Comparison of Literacy Model to the Framework Model	163
Figure 6.4 Literacy Model by Year level	164
Figure 6.5 Literacy Model by Year level and gender	165
Figure 6.6 Numeracy mean scores-Cross-sectional view with model trajectory	167
Figure 6.7 Numeracy mean scores –Longitudinal view with model trajectory	168
Figure 6.8 Comparison of Numeracy Model with the Framework Model	170

Figure 6.9 Numeracy Model by gender	171
Figure 6.10 Numeracy Model by Year level	172
Figure 6.11 Numeracy Model by Year level and gender	173
Figure 6.12 Summary of the Literacy Model and Numeracy Model by Year level and gender	174
Figure 7.1 English 1997 – Histograms of score distributions by Year level	180
Figure 7.2 Teacher Judgement assessments - English Learning Area 1997 by strand and Year level	182
Figure 7.3 Teacher Judgement assessments - English Learning Area 1997: means, medians, standard deviations and inter-quartile ranges, by Year level	183
Figure 7.4 Teacher Judgement assessments - English Learning Area 1997: Mean profile level of Reading and Writing strands combined, by age	183
Figure 7.5 Teacher Judgement assessments - English Learning Area 1997: Mean profile level of Reading and Writing strands combined, by gender of students	185
Figure 7.6 Teacher Judgement assessments - English Learning Area 1997: Mean profile level of Reading and Writing strands combined, by age within Year level	186
Figure 7.7 Mathematics 1998 – Histograms of score distributions by Year level	188
Figure 7.8 Teacher Judgement assessments- Mathematics Learning Area 1998 by strand and Year level	189
Figure 7.9 Teacher Judgement assessments - Mathematics Learning Area 1998: means, medians, standard deviations and inter-quartile ranges by Year level	190
Figure 7.10 Teacher Judgement assessments - Mathematics Learning Area 1998 Mean profile level all strands combined, by age	191
Figure 7.11 Teacher Judgement assessments- Mathematics Learning Area 1998 – Mean profile level of all strands combined, by gender of students	192
Figure 7.12 Teacher Judgement assessments - Mathematics Learning Area 1998: Mean profile level of all strands combined by age within Year level	193
Figure 8.1 Comparison of Equi-percentile equating by separate Year levels 3 and 5 with the combined data set for Years 3 and 5 - 1997 English.	204
Figure 8.2 1997 Profile to Test scale equating by equating method, Year 3 and 5 data combined- English	206
Figure 8.3 1997 English Teacher assessments – Conversion of Profile to Test Scale result: Independent Rasch model.	210
Figure 8.4 1998 Mathematics Teacher assessments – Conversion of Profile to Test Scale result: Independent Rasch model	211
Figure 8.5 A comparison of the final result of the unanchored conversion of the teacher scale to the test scale compared to the anchored result	212
Figure 8.6 1997 English/Literacy - Scatterplot of Teacher assessment and Test assessment invariance	213
Figure 8.7 1998 Mathematics/Numeracy - Scatterplot of Teacher Assessment and Test assessment invariance	215
Figure 8.8 Match rates 1997 - English/Literacy	217
Figure 8.9 Match rates 1998 - Mathematics/Numeracy	217

Figure 8.10 1997 English/Literacy: Comparison of Teacher assessments (Rasch model equated) and Test Model assessments at selected sites.	219
Figure 8.11 1998 Mathematics/Numeracy: Comparison of Teacher assessments (Rasch model equated) and Test Model assessments at selected sites	222
Figure 8.12 1997 English/Literacy-Mean of Teacher assessments (Rasch model equated) and Test Model assessments compared, by gender and Year level	224
Figure 8.13 1998 Mathematics/Numeracy-Mean of Teacher assessments (Rasch model equated) and Test Model assessments compared, by gender and Year level	225
Figure 8.14 1997 English/Literacy Test and Teacher mean scores at each Year level-Expression to equate means	227
Figure 8.15 1997 English/Literacy-Comparison of original teacher trajectory with the Year level mean re-scaled teacher trajectory and with the Test model trajectory	227
Figure 8.16 Effect of Alternative equating processes on Teacher Test assessment comparisons-using Mathematics/Numeracy 1998	228
Figure 8.17 1997 English-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age	230
Figure 8.18 1997 English-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by gender by age	231
Figure 8.19 1997 English-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age within Year level	231
Figure 8.20 Plots of points from Test and Teacher assessments from Figure 8.19 (Points are restricted to those within the appropriate range for each Year level)	232
Figure 8.21 1998 Mathematics-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age within Year level	233
Figure 8.22 1998 Mathematics Plots of points from Test and Teacher assessments from Figure 8.21 (Points are restricted to those within the appropriate range for each Year level)	234
Figure 8.23 1998 Mathematics-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age within Year level: Female students	235
Figure 8.24 1998 Mathematics - Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age within Year level: Male students	235