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I understand that as an unpublished work, a thesis has significant protection under the Copyright Act and I do not wish to place any restriction on access to this thesis.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Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

14/12/2010

Signature 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

List of Tables

List of Figures

Chapter 1: Mapping the topics of interest

> …when an author auditions his main character, he doesn't really know if he'll pull his weight in the novel until it's too late to choose another one.
>
> Hilary Mantel, 2005.

> *A vision for the future is that assessments at all levels—from classroom to state—will work together in a system that is comprehensive, coherent, and continuous*. In such a system, assessments would provide a variety of evidence to support educational decision making. Assessment at all levels would be linked back to the same underlying model of student learning and would provide indications of student growth over time.
>
> Pellegrino, Chudowsky & Glaser (Eds.), 2001, p. 9.

Overview of this chapter

The purpose of this first chapter is to introduce the main character, teacher judgement assessments, and to outline a number of topics deemed relevant to understanding assessment-managed learning. The thesis is an evidence-based 'thought experiment'. What if teacher judgement assessments could provide the critical data needed to optimise the learning growth of every student?

To focus the character development two propositions are proposed. One is that teacher judgement assessment is already of such quality that classroom, school and system assessments could be based on teacher judgement alone. The second takes a less radical position, proposing that there is sufficient evidence to support the notion that assessment based on teacher judgement has the potential to provide most of the data needed to improve the effectiveness of teaching and learning in schools.

To distinguish between these two propositions a range of issues relevant to teacher judgement as an instrument of student assessment and improved student learning are reviewed. This first chapter sets the general context. The propositions are detailed and the general questions to be addressed outlined. An understanding of how 'learning' is understood in this thesis is described. Approaches to the measurement of learning and diagnosis of the support required are outlined and consideration is given to the reasons why improvements in measuring learning within classrooms might improve student learning. The use of standardised and school system-wide tests as one approach to the measurement of learning, and as a reference for teacher judgement assessments is considered.

Assessments based on teacher judgement, whatever their current quality, are ubiquitous classroom practice. Processes to understand and enhance their quality, along with how

regular individual students assessments might increase the effectiveness and targeting of instruction, are considered.

The three Ps of Fullan, Hill and Crevola (2006) required for a 'breakthrough' in improving classroom instruction; Personalisation, Precision and Professional Learning, help frame this exploration of assessment-managed learning. To these a fourth P, 'Progressions' (Critical Learning Instruction Paths [CLIPs] in Fullan et al. terms) is added. These progressions, this thesis speculates, might provide teachers with more than just reference maps to assist the observation and management of learning. Progressions may help address a key problem with level structured curricula, inadequate processes for recording progress within a level, making the level structure very limited as a scale of learning progress.

More than a reference map however is required to assist teachers in easy understanding of what are the most effective options for support to students. Fullan et al. introduce the concept of a knowledge base (Fullan et al., 2006, p. 82). The knowledge base would provide access to relevant research and practical advice for teachers. Elements perceived to be relevant to the advocated knowledge base that would assist teachers in their assessments and the consequential management of learning are addressed in subsequent chapters.

Overview of subsequent chapters

Chapter 2 covers aspects of the early 20th century history of assessment. Scales for describing or recording learning were first developed in this period. The early scales (constructs) are examined to illustrate that their developers had already established techniques that might have enhanced the role of teachers as managers of individual learning through assessment, had these techniques developed differently.

Chapter 3 covers more recent curriculum and assessment developments of the 1980s and 1990s. It documents how the general principles of teacher judgment were adopted in the unrealised 1990s Australian national curriculum, and then in South Australia. It further outlines how the data analysed later in the thesis came into existence.

Chapter 4 reviews teacher judgement assessment and what studies have revealed about teachers' skills in estimating learning status. How judgements are made, how they are recorded and how well they compare with other forms of assessment are all addressed.

Chapter 5 illustrates how general learning trends are made more transparent by test data. Cross-sectional and longitudinal views of grouped and individual student data identify the understanding of learning that time series data provide. Elements of this analysis typify the information that might be part of the Fullan et al. advocated knowledge base. Models of test

data with age and Year level[1] are developed to estimate how test scores change with Year level. Test results for Year levels not tested can then be imputed from the models. Case studies, where information from psychometric analyses provide a refined appreciation of the challenges to students in learning specific skills, are identified as examples of ways to create observational tools for teachers.

Chapter 6 summarises data from tests for South Australia and estimates test data for Year levels where students were not actually tested. The findings from Chapter 5 relating to trajectories of learning are applied to support the development of the model to impute missing data.

Chapter 7 summarises teacher judgement assessment data for cohorts from Year levels 1 to 8. These assessments use the implied scale of each strand of the level curricula applying in South Australia in 1997 and 1998 in English and mathematics, to provide a level achieved for each student and the teacher's estimate of the student's progress towards meeting the criteria for the next level.

Chapter 8 draws the two perspectives together to establish how well one matches the other. The historical development, the teacher judgement review and the data analyses are then used to provide, in the concluding chapter, a basis for speculating about the role of teacher judgement assessment into the future.

Elaborating the main character: Teacher Judgement Assessment

Schools and classrooms are full of data but not often in forms that provide an understanding of individual student learning development. The prime data that should be of interest are those that will indicate how each student is progressing, how his or her learning is growing. For all the current data available, in the form of grades, marks or outcomes achieved, it is rare that data can be provided in a form that illustrates, particularly to an individual student, that the student's learning is growing.

An alternative source of data can be considered: the principal character of the thesis, 'assessment through holistic teacher judgement' regularly referred to in brief, as 'teacher judgement assessment'. Teacher judgement assessment is rather elderly, already over 100 years old. The history presented in Chapter 2 and the deeper analysis of some data from the 1990s raise some possibilities for using data from teacher assessments in a similar way to standardised test assessments, but allowing greater flexibility in how the data are obtained.

---

[1] Throughout 'Year level' is capitalised to avoid confusion with calendar year and/or (curriculum) level. On occasions Year level and Grade are used interchangeably.

It is argued that teacher judgement assessment is ubiquitous; it is entangled in all classroom assessments. In its usual current form teacher judgement assessment does not conform to standards that provide data adequate for students to self-reference their development or for classrooms and schools to have data for longitudinal purposes. The thesis is concerned with whether it is feasible to adapt teacher judgements in such a way that they can be considered to be consistent across teachers through linking the judgements to common criteria organised as development scales. Teacher judgements might then provide the student self-referenced data which can indicate, in general terms, skills[2] under development as a standardised test might, be easily recorded in an longitudinal student record system and be able to be used to make the learning visible.

Teacher judgement assessment presumes that a teacher, provided with appropriate background information about likely sequences of skill development and with the opportunity to observe students daily, using whatever observation tools they choose to apply (observation, conversations, teacher designed tests, standardised tests and myriad other possibilities), can posit a hypothesis about where each student is placed in their learning status. It is assumed that many teachers already hold conscious (or subconscious) hypotheses, but that the language for expressing these hypotheses is limited and ambiguous.

Under processes proposed later the teacher would refine any hypothesis about a student by integrating all the observations into a single judged learning status estimate for any given strand of learning and record the value in a database of student records. A strand is a cohesive set of skills within a learning area that can be seen as having developmental order and dependence relationships. Prior skills are required to be consolidated before later skills in the set are established. Based on the time-series of assessments for a student, the teacher would reconsider the form of support required through reviewing the trajectory of data points to that time point. Reviewing the trajectory could be as simple as the teacher looking at the graph of learning status over time. Reviewing might also draw upon more sophisticated analyses using learning models built from a range of statistical models drawing on teacher judgement and test data. These analyses might draw on artificial intelligence approaches to

---

[2] For convenience and ease of writing (and reading) the term ‘skills’ is used generically here to cover a wide range of similar and approximately similar terms such as knowledge, comprehension, skill, thinking strategy or behavioural disposition. While the distinctions in meaning are important in many circumstances, and might have significantly different connotations, for the purpose of most of the arguments in this thesis the use of the generic term will simplify the expression of the ideas. This process is consistent with the approach adopted by Rupp and Templin (2008) for a similar purpose in their review of diagnostic classification models (see Rupp & Templin, p. 228).

offer suggestions to teachers about particular students, an implied feature of the Fullan et al. knowledge base.

Thus the essence of the thesis is: Could teacher judgement data meet the requirements needed for the longitudinal recording of individual student learning? Can teacher judgement assessments provide the critical data needed to manage and optimise the learning growth of every student?

One technical approach to establishing learning status is the use of well-designed tests, aligned to the outcomes to be attained. Based on the Rasch model of test analysis, the scale of difficulty of test items provides a scale for measuring the learning status of the students. The unit of the Rasch scale is the logit, the log odds unit, with items spaced in terms of their relative difficulties.

This thesis considers the feasibility of using the *test scales* to report learning status, as distinct from requiring the use of specific tests. These scales might be able to be used more broadly for formative purposes. The scales provide a numerical language to record a student's learning status at any time. The value recorded has meaning in terms of skills likely to have been mastered and those under development. If test scores and the test scales to which they relate are to be seen as the currency of learning as implied by National Assessment Program Literacy and Numeracy (NAPLAN) in Australia and No Child Left Behind (NCLB) in the US, is it possible that teachers can use the same currency by using the test scale in their own assessment processes?[3]

This approach emphasises the construct rather than the particular assessment (Wiliam, 2010; Wright, 2001) and derives its usefulness from the positions of specific skills on the construct scale. It is the locations of these increasingly complex skills that then give meaning to the scale. The changing locations of students as their learning progresses along the scale indicate skill sets achieved and skills under development. The freeing of the construct from particular

[3] It is not critical at this stage to be concerned about which of a wide range of possible options for the numbering convention for a given scale might be used. The principle of teacher assessments and test assessments using the same scale convention is all that is required to be considered. The language of the *test scale* (the numerals and their meanings) could be used by teachers, or the language of the teacher scale could be used in reporting test scores. For the consideration of the principle either is possible. In practice the teacher scale has already been used to report test scores and teacher judgement assessments in Victoria and in the UK national curriculum. Should the reader consider that the feasibility of teachers using the test scale must have been established already by the existence these two examples, the published evidence is limited (covered in Chapter 4) and the use by teachers effectively limited to summative assessments.

tests and specific forms of assessment enables higher-level skills, unable to be assessed by bulk testing (deep understanding of an idea as an example) to be incorporated into the assessment scheme.

Wright (2001) argues the feasibility of a notional general construct, a scale of Academic Achievement Units, based of a broad combination of unspecified assessment processes whereby a scale of 1000 units (based on a transformation of the logit unit) might run from 0 for first grade to 1000 for college entry.

> Immediately test practitioners, teachers, parents and students know accurately where the student stands; how much the student has advanced; how much is yet to go; and the difficulty level of the material to teach next. (Wright, 2001, p 784)

This description is from the boldest of all the advocates of the Rasch Model. Such a unitary scale, assuming a single underlying dimension of learning, is less likely to be practical for individual teachers as might be separate scales for each broad aspect of the school curriculum e.g. strands as raised earlier. A more practical set of vertically aligned scales for specific skills is advocated by Jorgensen (2004).

Supporting the broadness of the possible approaches to assessment, Griffin (2007) explains that the Rasch model can be applied very generally, particularly the Linacre (2006) expression of the model, which has very few restrictions on scoring procedures. Accordingly the nature of the tasks used to establish the likely range of a student's skill development is very wide:

> The task could be a test question, a set of multiple choice items, an essay, a performance, a speech, a product, an artistic rendition, a folio, a driving test, the dismantling and reassembling of a motor car engine, building a brick wall, giving a haircut to a client, or whatever was related to some attribute of interest. The attribute could be an ability, an attitude, a physical performance, a procedure, an interest, a set of values or a generalised competence in an area of learning. (Griffin, 2007, p.89)

This description of the interaction of construct scales and approaches to assessing students highlights the importance of observation and the integration of expressed behaviours in the assessment of students at any time. The prime agent for doing this would seem to be the teacher.

Assuming it were possible for teachers to estimate student locations on the construct scales directly, a range of pedagogical options flow from knowing the current position. This thesis does not deal with the specific pedagogical consequences for any given location on the scale, this being the domain of teachers and a wider range of support experts. Teacher judged locations on the construct scales, however, provide a possible basis for the simple and regular

recording of the longitudinal progress of students, totally compatible with tests designed for the same constructs.

The keeping of meaningful records without wasting the time of teachers is one perceived benefit. Optimising teachers' professional roles using their observation and expert judgement skills is another benefit. The practicality of the process depends upon, among other factors, the extent to which teachers can make their judgements consistently, using the test construct scale and whether the test scale itself reflects an adequate model of the actual learning development. Alternative methods of recording learning progress using checklists of skills achieved or using rubrics within levels or sublevels may be adequate for some purposes but do not provide the utility of the construct score. The benefits of the scale approach to learning are developed throughout the thesis.

The essence of the Rasch model in test analysis is that student scores and test item difficulties can be aligned on the same scale. The student receives a score that has meaning in terms of what it estimates the student can do, give or take an error of estimation in the score. In principle, subsequent tests aligned to the same scale provide student scores with which earlier scores can be compared to reveal growth. The distance between scores indicates the amount of growth, subject once again to the impact of measurement error. Points along the scale have meaning in terms of indicating the extent to which specific skills are likely to have been developed.

The test process is however limited, expensive and constrained in the range of skills and behaviours that can be assessed. Even with increased frequency of tests, reduced time-lag in score provision and improved estimation processes through computer managed custom tests built in real time for each student, the amount of data provided to classrooms is likely to be small relative to that able to be provided directly by teachers.

Assessment as a support to learning

There is strong evidence that the use of assessment data can improve the effectiveness of teaching (Black & William, 1998; Crooks, 1988; Hattie & Timperley, 2007). The evidence also suggests that improved teacher effectiveness requires professional development in the understanding and use of assessment data (Timperley, 2009).

As considered above, the wide range of observations made by teachers provide a rich source for them to continuously hold hypotheses about the learning status of each student. These observations can be systematically planned - in the form of teacher developed tests, projects, assignments, probes, listening to reading aloud, student reports, standardised tests, conversations with individual students or any other planned observations. Also likely to play

a part in hypothesis development are unplanned observations of class events, casual conversations, general student interactions and other spontaneous behaviours.

Can a process be anticipated where teachers are assisted in the integration of their observations such that their observations, and those of colleague teachers, can be recorded in a consistent way?

This is considered as theoretically possible through the combined use of holistic teacher judgements, empirically developed learning progressions (average learning orders), across-teacher moderation and some simple processes to record the current hypothesis on learning status of each student. One option for simplifying recording, as raised earlier, is the scales of tests adapted for teacher assessment without the need to necessarily use the tests themselves. An anticipated result is a radically altered language for consistently documenting and monitoring learning.

These thoughts are not new. Fisher (1862) and Thorndike (1912) outlined views consistent with this approach and a range of approaches has developed since then. The current national curriculum in England, and the 1990s approach to a national curriculum framework in Australia, provide some ingredients for the thesis considerations. Teachers in England and Victoria have already been required to use teacher judgement assessments in parallel with test assessments. How well these assessments match tests assessments is revealed in Chapter 4. Teachers in Scotland and more recently in Wales, along with teachers in Queensland and the Australian Capital Territory, have used teacher judgement assessment (with moderation processes) for all assessment purposes including summative school graduation assessments. While these latter Australian examples are discussed briefly in Chapter 4 the thesis concentrates of cases where parallel assessment by teacher judgement and testing have occurred.

The works of Fullan, Hill and Crevola (2006), Griffin (2004, 2007), Forster and Masters (2004), Masters and Forster (1996), and Wilson (2004) on sequences of likely learning orders, leading to scaled learning progressions or maps, provide a basis for monitoring learning development. The judgement assessments of many teachers in the 1997 and 1998 in South Australia (see Chapters 7 and 8) provide an insight into whether teachers can judge the learning status of students with similar results to tests.

Teachers alone, as the prime resource for each student, have the potential to integrate and manage learning development. Unsurprisingly, effective schools research indicates it is the teacher that is the most critical resource in enhancing learning (Hattie, 2003; McKinsey & Company, 2007; Rowe & Hill, 1996). If teacher judgements were used to provide time ordered (longitudinal) learning data for each student, a rich basis for reflection on and

reconsideration of instructional or learning approaches would be available. In the manner similar to that of the medical professions, diagnosis and treatment would be integrated through their interaction over time. As for medical assessments some laboratory test data are available in the form of standardised, online and statewide test results. The integration of that data to resolve what to do next with each student is a professional judgement of each teacher. No other educational agent can both integrate the data and apply the appropriate treatment.

The unfolding of the issues in the thesis expands on what the general benefits might be. The longitudinal view of individual students should contribute to the *breakthrough* in instruction hoped for by Fullan et al. (2006). A breakthrough defined by Ellmore, in the foreword to Fullan et al., is "a sudden, dramatic and important discovery or development…[and/or] a significant …overcoming of a perceived obstacle, allowing the completion of a process." (Fullan et al., 2006. p. xi) Teacher judgement assessment with the right support is seen as potential key contributor to the breakthrough. In the process, the professional role of the teacher can be markedly enhanced.

Propositions considered

The thesis proposes two propositions to be examined.

The first proposition

The principal proposition is that teachers' judgements of students' learning status (scale values), in school systems where they have been applied, are valid indicators of student learning status for all students and for all teachers, and are already of such quality and reliability that classroom, school and system assessments can be based on teacher judgement alone.

The second proposition

The second proposition is that teacher judgement assessment can be enhanced to the point where it can provide valid indicators of student learning status, in the form of scale values.

As a metaphor of the general continua of learning, the first proposition is at the extreme positive end. The second proposition can be placed at varied positions on the continuum based on the evidence and speculation about potential. One possibility is the extreme of 'not feasible to enhance' (zero). Many other placements are possible.

Evidence to examine these propositions is obtained from the degree to which teacher judgements a) are internally consistent and b) can consistently match independent test assessments such as those obtained from statewide or national tests. As a methods

comparison problem, it is postulated that tests and teacher judgements are alternative methods of assessing the same learning dimensions.

Key questions considered

To explore the propositions a number of related questions are addressed regarding the relationships of tests and teacher judgement. These are:

1. What is the history of student assessment using scales to establish learning status? In particular what is the history of teachers using observer judgement?

2. What does the research literature on teacher judgement say about what teachers do and how well they do it?

3. What does analysis of the 1990s data from the South Australian adoption of national profiles (Curriculum Corporation, 1994a) reveal about the ability of teachers to estimate the position of students on scales described by increasingly complex learning behaviours?

4. What proportion of SA teachers were effective on-balance assessors of students?

5. What do teacher-generated and test-generated data reveal about the learning development of students throughout their 12 or more years at school?

6. Assuming some teachers are relatively effective on-balance assessors, what tools and processes might be required to maintain and enhance their skills and to develop the skills of less effective assessors?

7. How might the design of classroom and school processes be changed to optimise the use of teacher judgements?

8. What options might need to be considered for those teachers who have limited abilities in on balance judgement unimproved by practice?

9. What would be the implications of greater use of teacher judgement assessment to pre-service teacher education?

Evidence establishing the effectiveness of teacher judgement and thus which of the two propositions is supported is considered. Understanding the abilities of teachers as on-balance assessors leads to the consideration of practices to support teachers to develop and maintain the quality of their assessments. A number of other issues are considered. Can scales of teacher and test assessment be regarded as equal interval, an attribute that would be fundamental to the application of statistical and arithmetic processes to student scores? How can teacher-generated data be used to track individual student development over time and across year levels? How might subsets of these data be used for school management

purposes? What policies and processes might be developed to take advantage of teachers' judgement skills?

An important consequence of the acceptance of either of the two teacher judgement assessment propositions will be the enhanced recognition of teachers as professional, trustworthy managers of learning. A negative finding would have significant implications for all assessment and teaching polices independent of method.

The scope of the topic is very broad. As a result a wide range of sources have been scanned for relevance. The task of selecting those to explore and cite has been difficult and clearly many sources and examples, deemed relevant in the view of any particular reader, may have been ignored. It is a reflection on the range of material potentially available that there are some bodies of work, ostensibly on the same issues, that do not reference each other. The treatment in this thesis makes no claims to being inclusive of all possible sources.

Learning: an operational definition for this thesis

Fundamental to this study is an operational definition of learning. Dictionary definitions are brief and inadequate. Example dictionary definitions are:

> Knowledge got by study. (Concise Oxford, 5th Edition)
>
> The cognitive process of acquiring skill or knowledge. (Princeton University WordNet 3.0, 2006)
>
> Knowledge or skill gained through schooling or study. (The American Heritage Stedman's Medical Dictionary)
>
> The act, process, or experience of gaining knowledge or skill. (The American Heritage, Stedman's Medical Dictionary)

None of these definitions is adequate in conveying the full understanding of learning as applied in this thesis.

Even the comprehensive *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001) in its recent significant treatment of learning, cognition and assessment, addresses learning without a clear initial definition of the concept, notwithstanding the recognised need to describe other important terms; cognition, cognitive sciences, educational measurement, assessment, and testing are all defined. An understanding of the use of the term is developed over many pages in that text. Based on a text analysis of cases of the use of 'learn' it can be inferred that learning as understood by *Knowing What Students Know* has some of the following elements.

Learning is a process that leads to a "transformation of naive understanding into more complete and accurate comprehension (p. 4). Learners construct their "understanding by trying to connect new information with their prior knowledge" (p. 62). Learning leads to "increasingly well-structured and qualitatively different organizations" (p. 71) as knowledge develops. Development and learning are differentiated. Some forms of knowledge are acquired by all or most individuals "in the course of normal development, while other types are learned only with the intervention of deliberate teaching" (p. 80). Studies of learning "show that in a responsive social setting, learners can adopt the criteria for competence they see in others and then use this information to judge and perfect the adequacy of their own performance" (p. 89). Expertise is developed by "practice and feedback" (p. 91). Learning is not just "a matter of acquiring more knowledge and skills, but as progressing toward higher levels of competence as new knowledge is linked to existing knowledge" (p. 115). Deeper understandings replace earlier understandings and "ordered levels of understanding and direction are fundamental: in any given area, it is assumed that learning can be described and mapped as progress in the direction of qualitatively richer knowledge, higher-order skills, and deeper understandings" (p. 115).

The understanding of the concept of learning adopted in this thesis is consistent with the elements above except in one respect. Apparent developmental (maturational) changes in performance are included as part of the understanding of learning. This is done for a practical reason, the inability in this study and in educational practice generally, to partition them out. Explorations of time related effects on student learning in school (Cahan & Davis, 1987; Hattie, 1999; Kissane, 1982) identify an underlying small but important effect of the passage of time (or small increases in age) on learning improvement.

Within-Year level effects are shown with age (Cahan & Davis, 1987; Grissom, 2004; Tourangeau, Nord, Lê, Pollack, & Atkins-Burnett, 2006; Williams, Wo & Lewis, 2007). For each 0.1 of a year increase in the mean age of cohorts grouped in these fine age categories, the mean scores on tests increase consistently until the mean age exceeds the normal age range for the cohort for that grade or Year level (see Chapter 5). For some subsets of students (the lower socio-economic status groups) the effect of a period of no instruction (summer break) is to go backwards (Alexander, Entwisle & Olson, 2001; Cooper, Nye, Charlton, Lindsay & Greathouse, 1996). However for many students the positive age/time learning trend appears to be maintained with time (Cahan & Davis, 1987; Tourangeau, Nord, Lê, Pollack, & Atkins-Burnett, 2006 shown later in Chapter 5, Figure 5.9). There are indications

in Adult Literacy surveys[4] that this positive maturational element combined with experience and ongoing skill refinement continues for many people into their early 30s, well after they cease formal institutional learning. After age 35 mean literacy and numeracy skills appear to reduce slowly with age, signifying an average negative impact of maturation from this age.

Thus reading or mathematics performance, on average, can improve for many students with time even when instruction is not occurring. The cause may be practice or maturation. Assessments of learning improvement over time for individuals or groups may therefore include a maturational element and it makes sense to make this clear in the definition of learning. In assessments related to scales of learning, analogous to rulers as considered in this thesis, the apparent maturational contribution to measured learning, whatever its cause, will be included automatically and unidentifiably in the assessments. An analogy with height is that the height measurement is not discounted for interactions between genes, maturation and nutrition. Together they lead to particular heights and rates of growth at phases of an individual's development. The inclusion of maturation (development in some descriptions) is not to imply a direct genetic influence in the differential learning rates of individuals, only that across individuals similar general trends with age/time apply.

Thus learning is defined for the purpose of this research to be an increase in knowledge, comprehension, skill, thinking strategy or behavioural disposition (generically called 'skills'), through experience, direct study or through some natural developmental change in cognitive functioning. The increase is seen as having greater complexity than just acquiring more skills, but rather, moving through ordered levels of understanding, progressing in the direction of qualitatively richer knowledge, higher-order skills, and deeper understandings.

The definition of learning proposed is strongly influenced by the vision of Wilson and Sloane (2000) and other advocates of the Rasch model for measurement. The proposed definition implies learning is an increase in the repertoire of behaviours. The increase can only be inferred from some externalisation of the behaviour or performance of the learner. What led to the increase and how it is managed within the mind is unknowable from external, non-intrusive, observation. Conclusions about causes for various rates of learning growth (treatments, forms of teacher intervention, maturation), while often plausible, are also ambiguous, inferential and probabilistic.

---

[4] Appendix 2 shows simple analyses of Australian and US data that suggest the literacy skills of the average population increase generally until age 30 to 35, even though only small proportions remain involved with formal education programs in schools, TAFE and universities.

It seems reasonable to assume that learning is most likely stored through a change in brain function but that understanding the storage mechanisms is not necessary for the observation and measurement of learning from a teacher's perspective. This is not to imply that the understanding of cognitive / memory processes is not useful for teachers, only that learning should be observable without an understanding of detailed brain function, provided the teacher is aware of what behaviours to be looking for.

Learning, whether active, induced or maturative, cannot be disaggregated readily into these component contributions, certainly not in the classroom. Experience (somewhat passive accommodation of external activity) or study (more active participation and interaction with information and individuals) are indistinguishable in their relative contribution to a changed state. They are not mutually exclusive categories. Whether any learning has occurred must be inferred through some external manifestation of the internal state of the learner, making knowing that state difficult.

In summary, learning is a dynamic process of the individual, for which a state value can be estimated at any point in time. A reading of that state can be taken by a specific standard interaction (some form of standardised pencil and paper or computerised test) or through a series of observations by, and interactions with, a teacher; or even self-assessed by the student. It is assumed that multiple processes can be used to estimate the quantum of learning at any time and that varied processes will arrive at essentially the same result. To know if the results by different processes are essentially the same requires the use of a scale that is common to all processes, or a process of transformation between scales such as that in Fahrenheit to Celsius temperature scale conversions. This requirement must be met whether different assessment processes are applied to an individual at a single point in time, or the same processes are applied over different points of time.

Progression

Complementary to the concept of a scale to quantify learning growth is the understanding of the likely order of development of particular skills and higher-level behaviours. One concept that comes out of the empirical analysis of learning growth is the progress map (Forster & Masters, 2004). The progress map is also known as a Critical Learning Instruction Path (CLIP) (Fullan et al., 2006) or more generally a learning progression (Popham, 2007; Heritage, 2008). These progressions can be traced back, in a form, to the handwriting and prose scales of Thorndike (1910) and Hillegas (1912) or even to Fisher (1862) discussed in depth in Chapter 2. The exploration of item orders in a graphical form by Thurstone (1925) provides another view of progression. Progressions offer a context for teachers as they make

personalised assessments. In principle they should help teachers make decisions about what support is appropriate for the student.

The Statements and Profiles for Australian Schools (SPFAS) (Curriculum Corporation, 1994a) approximated a form of progression. The profiles described criteria for achieving each particular level within any strand within a designated learning area. The specific criteria for a level, however, did not have a likely order of achievement. For this reason it was hard for teachers to record the progress of learning within a level. The teacher could report that some, all or most of the criteria had been met. Had some criteria been provided with empirically confirmed probabilities of being met earlier than others, the criteria themselves could have been seen as spaced along the dimension as useful indicators of progress.

Progress maps have been refined through applying Rasch model approaches to place skills, as distinct from test items, in empirically derived orders. The maps/scales are used in some school systems, through teacher observations, to support classroom based teacher-assessments and curriculum development. The classroom applications take the form of ordered descriptions, appropriately spaced, of what students can do at particular points on a described spectrum of tasks, skills and items (Forster & Masters, 2004; Heritage, 2008; Popham, 2007).

These scales (or maps) are, in principle, independent of the particular testing or observation practices that have led to their creation. That is, a variety of methods can be used to estimate the learning status for any student. The scales illustrate a most likely order in which understanding (or learning) develops. As well, what might be demonstrated by a student at a particular point on the spectrum, and the relative learning distance between skills, can be described. The order is quite likely to represent a dependency relationship between successive skills. The learning distance is an estimate of the relative difficulty of any particular skill compared with an easier skill. Learning distance is then a likely correlate of time to learn as well as the probability of success. The Rasch model can be seen as a tool to assist in the scaling of a set of ordered skills in a way that might assist teachers in monitoring how a student is progressing. Progress maps are taken up in later chapters as one technique to assist assessment precision and to help understand progress in learning.

The map or progression need not be an ordering of particular skills but an ordering of tasks of known difficulty. Learning progressions equivalents for reading can be created by the use of tools that establish a difficulty level for a text. The Lexile Framework (Stenner, Burdick, Sanford & Burdick, 2007; Stenner & Stone, 2004) uses the Rasch model to establish the difficulty of texts on the basis of word frequency and sentence length. Observing the quality of the interaction of a student with a text of known difficulty provides a basis for a teacher to estimate the reading learning status. A complementary tool to estimate the difficulty of

mathematics tasks (Quantile Framework for Mathematics, 2010) provides a parallel process in mathematics that could be used for direct teacher estimation of learning status. Both processes provide the equivalent of learning progressions based on task difficulty, to assist teachers in their observation and management of learning.

More than progressions, however, are required to assist teachers in an understanding of what are the most effective options for support to students. Fullan et al. (2006) introduce the concept of a knowledge base of "integrated … expert instructional systems in which the hard work is taken out of the task of collecting and using the data." (p. 82). Among a number of elements, the knowledge base would include links to "updated information on students, their progress and other relevant classroom, school and system characteristics." (p. 82). The proposition of this thesis is that much of the progress data may be able to be obtained from teacher judgement assessments. Held in the knowledge base would be the progressions (CLIPs), advice on teaching strategies, appropriate resources, research evidence and a range of other information useful to the teacher. Access to relevant research and practical advice would be at a teacher's fingertips along with the opportunity for teachers to report their own success with particular intervention strategies. Elements perceived to be relevant to the proposed knowledge base are addressed in subsequent chapters.

Personalised learning

Fullan et al. (2006) argue that instruction should be personalised, as should assessment. For assessment to be personal it must be seen as a personal event rather than a group event. The student needs to perceive the teacher as considering and responding to him or her alone. This needs the teacher to show interest in the student personally through conversations, sensitive questioning and taking account of personal products and behaviours. To make a personal approach practical the teacher needs techniques for assessment and recording that are easily carried out. One element that might influence that ease is the map for the general journey and the progress the students makes along the likely learning continuum for particular curriculum areas. The more understanding the teachers has of what to expect, based on the general empirical observation of student development, the more prepared the teacher is for the task of relating the personal to the likely patterns of development. The better understood the framework, the easier it is to see where each student seems to fit at any time. Insights into the richer context for learning development are available to teachers from empirical research on student learning. The general patterns from system wide and international testing provide some of those insights but the path of each student is likely to be unique and not necessarily related to the group patterns. The complex issue of individual growth is addressed briefly in Chapter 5.

Understanding learning with student test data

Data from the application of tests to large cohorts of students have helped researchers understand some features of learning in school populations. It is unlikely that teachers, in general, are aware of the subtleties of these findings relating to learning within Year level by age, rates of learning development and gender patterns. Thus it is unlikely that teachers have the knowledge and ability to manipulate their assessments based on what is known to be obtained from tests, so that their assessments will reflect the same subtleties by intention. If teacher assessments of students do reflect these subtleties it is likely to be confirming of the validity of teachers observations. Chapters 5 and 6 provide an understanding of what test data show about rates of improvement with Year level and by age within Year level and learning area specific gender patterns. Chapter 8 explores whether teacher assessments show the same patterns.

Strengths and limitations of the study

One of the strengths of the study is its uniqueness. Assessment data from South Australian teachers in 1997 and 1998 are compared with tests for the same student populations. The author is unaware of any comparable data set that provides insights into the skill of teachers' on-balance judgements of student learning status when compared to test measures, although similar data may be available in Victoria and England. Furthermore the number of cases included is large and a number of convenient replications are included (two calendar years, two grades of actual test data, 8 Year levels of teacher assessments). The data provide insights into the skills of teachers as assessors using a curriculum framework for making their assessments.

One limitation of the study is the lack of information about multiple judgements from individual teachers. It was a deliberate design requirement of the data collection that teachers not be identified. A few cases of anonymous individual teacher patterns can be established in smaller schools, where only one class in a Year level is offered. Since the assessments cannot be grouped by individual teacher and because of the low number of cases per teacher, it is not possible to establish the variability within the assessments of individual teachers with confidence. This makes the size of the professional development task difficult to estimate for policy purposes, if say the second proposition of the thesis is accepted.

A further limitation is the inability of the study to track individual students over an extended period, to see how teacher judgements (or test scores) track learning status with time. The literature review considers a small number of cases where the variability in the patterns of individual student development is considered. These longitudinal studies make clear the wide range of patterns likely to be encountered by teachers observing the learning growth of

students. The data analyses in Chapters 6, 7 and 8 provide no information on these individual student learning trends. An implication of the consideration here is that all teachers should be seen as longitudinal researchers using regular on balance judgements through observation and other techniques including standardised testing where feasible, to document and manage the learning of individual students.

Summary

This thesis is a speculative search addressing what might be required to adjust assessment practices in classrooms so that student-learning growth can be made visible to, and by, teachers. Specifically, the ways in which the observations of teachers can be converted into scale values representing student learning status are considered. The thesis is concerned with examining possibilities rather than proving or disproving the validity of new ways to manage student learning.

The data analyses are important in providing evidence to judge the relative acceptability of the two main propositions. However, the bigger picture issue of how calibrated teacher judgement assessments might be developed to underpin student assessment processes, and thus the learning processes in schools and school systems, is of prime concern. Teacher judgement assessments, whatever their current quality, are ubiquitous classroom practice. Thus there is a need to make them of as a high a quality as possible.

The three Ps of Fullan et al. (2006), Personalisation, Precision and Professional Learning, required for a breakthrough in classroom instruction to a "more precise, validated, data-driven expert activity that can respond to the learning needs of individual students" (Fullan et al., p. xv), help set the scene for framing the exploration of assessment-managed learning in the classroom. To these a fourth P, Progressions (Critical Learning Instruction Paths in Fullan et al. terms), can be added. These progressions, it is speculated, might provide teachers with a reference map to assist the observation and management of learning and address a key problem of learning scales within level curricula, the lack of detail for progress within a level. As a consequence of the wide gradation increments inherent in level structures, teachers' judgement assessments are restricted to a small range of values. Alternative scaling processes might allow a range of values where scale increments could relate to days or weeks of learning rather than to months.

Summaries of changes in the means of learning status as Year level and age increase are considered through the research literature and particular school system records. This research assists in the development of a hypothetical model to estimate test data for untested SA Year levels as part of the data analyses. It also establishes patterns of cohort development as shown in longitudinal studies. Since the trajectories of individual learning appear to vary

markedly from the trajectories of the means of cohorts, individual pathways of learning development over time are addressed briefly. The degree of comparability of teacher and test assessments is then described, drawing on data from the South Australia school system. The final chapter draws together the key findings of the thesis into general conclusions about the acceptability of the main propositions and addresses the questions raised earlier in this introductory chapter.

The issue of quality and consistency in teacher judgement assessment is not new. The next chapter establishes that approaches to the consistency of classroom teachers' assessment have been considered for more than a century. Some of these approaches had the unrealised potential to build on teachers' judgements as the prime source of consistent classroom derived data documenting student learning.

Chapter 2: Early approaches to quantification of learning and scale development

> It seems that too often that we, and students in particular, are remiss in studying the history of our field. This is unfortunate because this historical background provides the framework within which we can interpret the value of current work as well as allowing us to assess the progress of our science.
>
> De Ayala, 2008, p. 209

The history of the early development of educational assessment is important because it reveals some of the thinking behind early attempts to make teacher judgement assessments consistent. One approach to the quantification of learning sought to identify examples of student work of increasing quality and to allocate values to the examples. These examples and their values were used as references for teachers to judge the quality of other cases of student work. Other approaches to quantifying learning required direct responses from students to test items. In both forms of assessment, the early developers created scales to order and space the quality of the work or the abilities of students. The scaling processes in these early developments indicate that the potential was there to better integrate the role of the teacher as assessor into the teaching/learning cycle than ultimately occurred. Two early innovators, Thorndike and Courtis, feature most significantly in the chapter and interact with a number of the other contributors, many of whom were students of Thorndike.

The chapter considers how information about learning in schools was obtained by the first researchers. Initial student examinations and surveys, and the instruments and processes adopted to carry them out, provide a context for the developments in the first 25 years of the twentieth century. As is the case of other scientific endeavours the earlier works inspired those who came next to refine and develop those ideas. The chapter describes the processes for obtaining information, and insights from early educational surveys and early approaches to creating scales, that led to the quantification of learning.

Placing examples or students on a scale required units for those scales. These units represented the value of the example or the score of the student, thereby providing quantitative values for learning. Thorndike, and later Thurstone, developed approaches to this. Interestingly, in hindsight, their solutions can be shown to bear a linear relationship to the log odds unit, adopted now as the logit of modern Item Response Theory and the Rasch model. In a consideration of teacher judgement assessment, this thesis argues that units are required for teachers to indicate learning status. The early history of quantifying learning provides examples that link the initial steps to improve teacher assessment consistency to the

potentially broader application of teacher judgement assessment as the prime source of learning information about students.

Timeline of the key examples considered

A number of case studies of the 19th and early 20th century illustrate the development of approaches to understanding classroom learning processes that were taken by administrators, teachers or researchers. Initially these take the form of examinations and surveys. New processes evolved from these into standards-referenced examples, drawing their examples from authentic student work. Mixed in with these developments is what are recognised today as pencil and paper tests. Table 2.1 below summarises the key periods and developments tracked in this chapter.

Table 2.1 Timeline of historical developments in assessment considered

Year	Development
1845	Boston: common exam on one day, holistic judgement of writing.
1850-1862	Greenwich: Rev. Fisher employs his scale book reference system for a range of subjects.
1892-1893	US: Rice's first survey of schools; mainly observational.
1897	US: Rice's survey of spelling; development of systematic standard approaches to data collection.
1902	US: Rice's survey of arithmetic.
1908	Stone, influenced by Rice's surveys, develops two arithmetic tests, 'fundamentals' and 'reasoning' to survey arithmetic in a number of school systems.
1909 -1911	Courtis replicates Stone's survey in his own school and then refines and expands the general approach to better understand improvement across grades. Courtis's tests become popular and provide comparison data back to Courtis for reference by teachers in their classroom use of tests.
1910	Thorndike develops the Handwriting scale.
1912	Hillegas (student of Thorndike) develops a scaled (holistic) judgment approach for the quality of prose composition Thorndike argues benefits of scale positioned standards as approaches to measurement in education.
1913	Thorndike adds items to the Hillegas (prose quality) scale.
1914	Ballou develops Harvard-Newton Scales; an alternative to the Hillegas scale. A set of instruments for composition, one for each of the four discourses. Thorndike (with Gray) develops a reading ability scale ('scale a' for visual vocabulary for single words) and a reading comprehension scale ('scale Alpha' for measuring the understanding of sentences). Courtis develops a composition scale similar to that of Rice.

Year	Development
1916	Hudelson advocates a standard running from 10 to 120 in degrees of difficulty; with the minimum for the first grade being 10, second 20, up to 120 for the twelfth year. Appears to be first conception of learning into *levels*. Trabue Completion test: an indicator of the ability to think about words and language. Items are scaled, based on Thorndike's approach and four parallel tests developed. Used as pre- post-tests to gauge annual progress.
1916-1923	Various new scales developed.
1917	Trabue adds items to the Hillegas scale (the Nassau County Supplement).
1925	Thurstone proposes approach to item scaling and visual representation; the first item map.
1950s	Rasch addresses ways of connecting test data over time and develops sample - independent establishment of item difficulty. Leads to new approaches to item scaling, test data analysis and scaling units.
(1984)	Engelhard, using Trabue's 1916 data, establishes that early approaches to scaling and item invariance by Thorndike and Thurstone approximate the Rasch approach.

1845 Massachusetts: the first system wide examination process in the US

Systematic approaches to examinations of students in a standardised form in the US are traced back to Massachusetts and are reported by Mann (Mann, 1845, reprinted in full in Caldwell & Courtis, 1925). While historical references emphasise the importance of Mann (e.g. Butts, 1978; Johnson, Dupuis, Musial, Hall, & Gollnick, 2002), the actual credit for the reported improvement in examination approaches goes to a close confidant of Mann, Samuel G. Howe (Good, 1926), a member of the School Survey Committees of Boston. These Committees, one for Grammar Schools and a second for the Writing Schools, had responsibility for reporting annually on each school, somewhat akin to one of the roles of Her Majesty's Inspectors in 19th Century Britain, although in Boston these were unpaid committee members. Sub-committees developed reports following one-day visits (Caldwell & Courtis, 1925, p. 195).

Prior to 1845 inspections had relied on oral tests and observations and were perfunctorily reported. In 1845 the procedures for the survey committees for reporting on each of the Grammar Schools and Writing Schools were radically modified to include a written examination of students (Caldwell & Courtis, 1925, p. 26).

A survey committee of three was established to report on Grammar Schools. The committee made clear its reasons for the written examination approach. Independence of process and an evidence base were seen as fair in reporting on the schools, though implied in the process was a mistrust of school staff. The committee applied assessment processes to give the same advantages to all (avoiding leading questions) so as

> to ascertain with certainty, what the scholars did not know, as well as what they did know; to test their readiness at expressing their ideas upon paper; to have positive and undeniable evidence of their ability or inability to construct sentences grammatically, to punctuate them, and to spell the words. (Committee Report, 1845 cited in Caldwell & Courtis, 1925, p. 26)

Historically, this represents a key change in assessment practice. The logistics adopted to achieve the fairness objective were comprehensive. Mann (1845) acknowledged parallel developments in written examinations in Europe and Great Britain. The processes for reporting on the quality of schools at that time in Europe appear much less comprehensive than those of the Boston committees. Reports on the processes of Her Majesty's Inspectors of the time suggest a model closer to the pre-revision Boston model (Arnold, 1889; Sneyd-Kynnersley, 1913; Wyatt, 1917). The examination boards of Great Britain, for example, were not established until 1857 for Oxford (Oxford University Archives) and 1858 for Cambridge (Cambridge Assessment Archives).

The Boston School Committees' reports document the beginning elements of school system wide approaches to examinations. Concepts of common test items, external control of timing, the time allowed for the test, security, approaches to analysis and a public report on the results were precursors to general system-wide testing as it has evolved today. The process did not provide assessment or pedagogical support to the classroom but was sophisticated for its time. It sets the scene for the external survey testing approach as a way of understanding the quality of schooling through a standardised assessment of the quality of student work.

1850-1862 Fisher Scale Book and numerical approach

The next regularly referenced innovation in quantified assessment is Reverend Fisher's Scale Book. Procedures to manage the quality of student work within individual schools, in the middle 19th century, are not well described in the literature of the time. One exception is a brief paper written by Reverend Fisher, Principal of the Greenwich Hospital School in the U.K. *c.*1862. Fisher was encouraged by Chadwick, President of the statistical section of the British Association for the Advancement of Science to document his assessment processes. Fisher's paper subsequently became a widely referenced 19th century example of an approach to teaching quality, student assessment and scale development (Ayres, 1918, cited in Cadenhead & Robinson, 1987; Cadenhead & Robinson, 1987; Haertel & Herman, 2005).

Fisher had been a naval officer, chaplain and astronomer in the 1820s and 30s, accompanying the Buchan and Parry expeditions to the Arctic in a role similar to that of Darwin and Huxley in their own scientifically formative voyages (Darwin, 1860; Huxley, 1936). Fisher's systematic scientific observational background might help to account for his quantitative approach to classifying the quality of student work.

Fisher presented his process in a paper to the 32nd Meeting of the British Association for the Advancement of Science in October 1862 (Fisher, 1862) with the title *On the Numerical Mode of Estimating Educational Qualifications, as pursued at the Greenwich Hospital School*. Fisher recognized the utility of a numerical representation. It was not only an efficient and information-rich mode of recording "easily referred to at a future time" but it afforded "also the means of determining the average condition of a class or school, as regards each subject of instruction and also the whole amount of educational work done." (Fisher, 1862, reprinted in full in Cadenhead & Robinson, 1987, p. 17). He was able to adapt the data for graphical representation, plotting the "mean values of the various educational qualifications of the boys at the completion of their education from 1850 to 1862 at each quarterly examination" (p. 17).

Chadwick (1864) reported an interview with Fisher who explains the rationale for the development of the arrangements put in place at Greenwich. "We had no records of results and it was to supply the deficiency that the numerical method was devised by me. The teaching was of a very inferior character." (Fisher in Chadwick, 1864 p. 263) Fisher developed his Standards Scale book in response to the perceived inadequacy of descriptive terms good, bad, indifferent and so forth, which he saw as subject to "various and somewhat uncertain interpretations, … arising from the fact that no recognized standard or fixed scale has hitherto been employed in assigning the absolute and comparative values of such expressions" (Fisher, 1862, reprinted in Cadenhead & Robinson, 1987. p. 16). In his own words his intention was to "refer such elementary attainments to standards which approximate to a permanent character to numerical equivalents for such terms, to afford more accurate and precise meanings than the words allude to, and at the same time [provide] a more concise mode of registration, combined with the means of integrating or expressing the sum-total of any number of results." (p. 16). The method used numerals one to five to denote the standard of work.

The scale-book contained examples of varying degrees of proficiency with a numerical value for each. For example, to "determine the numerical equivalent to any specimen of writing, a comparison is made with various standard specimens of writing contained in this book, which are arrayed and numerically valued according to the degree of merit" (Fisher, 1862 in Cadenhead & Robinson, 1987. p. 16). This anticipated by 50 years a similar process developed by Thorndike in 1910. The scale's highest value was 1, lowest 5. Scale points at a

quarter of a division were used to denote intermediate values, allowing in all 17 scale positions from 5 to 1[5].

The scale-book also included spelling, mathematics, navigation, scripture knowledge, French, general history, chart drawing and practical science; in these cases providing questions in each subject, to serve as types of the difficulty and also the nature of examinations. The scale-book did not include reading, characters and natural talents "where the usually received interpretations of the words 'good', 'bad' etc." (Fisher, 1862 in Cadenhead & Robinson, 1987. p. 17) were deemed adequate. Rev. Fisher did not regard reading performance as needing to be scaled.

The system had utility in the Greenwich Hospital School but seems not to have spread too far into the English school system. How the teachers in the school might have felt about the process is not reported. The scale book was systematic and numerical and is preserved for posterity through Chadwick's interest in a quantitative approach to documenting education.

1892-1908 Rice's educational surveys and Stone's enhancements

In the U.S. the next widely acknowledged developments in the evolution of educational assessment are the surveys of Rice. His initial survey of 1892-93 was observational, richly described and included student work collected from his visits (Rice, 1893). He observed the general instruction independent of topic, looking in particular for 'scientific teaching'. His methodology changed for subsequent surveys in spelling in 1897 and arithmetic in 1902 (Rice, 1913). In these latter surveys he collected written responses and provided statistical analyses. His findings confounded the administrators and teachers of the day. Schools providing 15 or 20 minutes of spelling daily did as well as those providing 40 or 50 (Rice cited by Thorndike, 1914a, p. 293). While Thorndike criticized the methodology (lack of recognition of the importance of the difficulty of words chosen to be tested) by implication he accepted the general drift of Rice's findings and acknowledged the importance of Rice's data collection approach (Thorndike, 1916a, p.5 and p. 9).

---

[5] Scale note: It is likely that Fisher assumed an equal interval scale. This thesis considers the utility of the logit (log odds unit) as a unit of measurement for learning development. Using assumptions about skill development over an extended period of time (5-7 years) and based on parameters from current test measures (Hungi, 2003) it is possible to estimate that a scale increment of a quarter of a scale unit ($1/17^{th}$ of the full scale) was likely to have been of the order of 0.2 logits in current terms (Estimated 3.5 logits from value 5 to value 1, divide by 17 units). Whether this level of precision is practical is an issue addressed later.

Stone, following a survey approach similar to that of Rice, collected data about arithmetic in 1908. He developed two arithmetic tests, one addressing 'fundamentals', that is the four operations of addition, subtraction, multiplication and division; the other addressing 'reasoning', the solving of word problems with relatively complex arithmetic and logic.

He acknowledged his debt to Rice (Stone, 1908, p. 96) but believed he had made improvements on Rice's approach. These were improved instrument and research design and improved methods of securing and handling data. The improvements in the gathering and handling of data were "chiefly those of refinement, and they could hardly have been planned for without the benefit of Dr. Rice's and other pioneer studies.' (Stone, 1908, p. 96) Stone was of the view that reasoning and fundamentals were different abilities "and should be so measured" (1908, p. 96). A number of other refinements were made to address, amongst other matters: time allowances for the test, test-room procedures, scoring, access to the data and computations (i.e. data analysis) by other researchers, disassembling the fundamentals into the four basic arithmetic operations (addition, subtraction, multiplication and division) and the use of correlation coefficients (Stone, 1908).

Stone's analysis was comprehensive, considering that all the tabulations and computations were completed by hand. He presented his data as aggregations of scores or aggregations of errors made. His major conclusion was: "Probably the truest single expression of the findings of this study is summed up in the one word, diversity." (Stone, 1908, p. 90) He noted that the within-system variability was greater than the between-system variability. In his view the greatest need identified by the research was the promulgation of standards of achievement. This need is understood today as a need for benchmarks. "That the great variability herein shown would exist if school authorities possessed adequate means of measuring products is inconceivable." (Stone, 1908, p. 90)

Of the surveys published to this time, Stone's was the first to show an appreciation of the issue of item difficulty. Stone's concern with item difficulty was twofold; he wanted to present items in tests in order of increasing difficulty and also to assign a weighting for more difficult items in the data analysis. He recognised the importance to his analysis of weighting the more difficult items. In this he was possibly influenced by Thorndike who gave "guidance in executing the statistical phases" (Stone, 1908, p. 5). He considered two options for weights. One related the proportion successful on the most difficult item (12% correct) relative to other items including the easiest item (94% correct), generating a wide range of weights with a direct relationship to the log of odds ratios (author analysis). His second option restricted the range to a maximum of 2 for the hardest items. As a result the weights for a selected a set of items were unity (i.e. weighting=1 for the easiest set), a hard set with a

maximum weight (2) and an intermediate set of 3 items with weights scaled at independent values between 1 and 2, on what can be established as a log odds ratio basis (author analysis).

Stone applied the weighted transformation to the total score for a system rather than to each student. As a consequence of the weighting, school systems had their aggregate score (items correct per 100 students) weighted in very approximate proportion to the log-odds transformation of the more difficult items (author established not detailed here). The resulting score per system was stretched for higher performing systems relative to lower performing systems[6], very crudely applying one principle of the logistic transformation used in the Rasch model. For its time this was an important and prescient insight by Stone but he had no reason to see the scale of his reasoning items as having a value beyond its contribution to stretching the score of high performing systems relative to lesser performing systems. His work, however, encouraged others to attend to what learning was actually occurring in schools.

1909-1911 Courtis and the influence of Stone

In 1908 Courtis was head of Science and Mathematic Department of the Liggett School, Detroit. Immediately after the publication of Stone's analysis he applied the Arithmetic tests to all students in the school (Courtis, 1909a). He published the results of a series of tests and the subsequent refinements he made to the testing process, in instalments, in *The Elementary School Teacher* (Courtis, 1909a, 1909b, 1910, 1911a, 1911b, 1911c). His first instalment applied the test unchanged but varied the scoring process. He did not adopt the weighting system applied in the reasoning test, thus moving away from the Stone insight of increasing the distance of higher scores from lower scores.

Initially Courtis's interest was in seeing how his students compared with Stone's data but, unlike Stone who tested only grade 6A, he was also interested to see the effect across the whole school. Thus he applied the test, unchanged, to Grades 3 through 13, testing 218 students in all. He described his reason as establishing standards for judging the success of a reorganisation of the mathematics course in the school. He also declared an interest in tracking

> … the development of ability in arithmetic from the primary grades through the high school. Such tests, repeated at frequent intervals, would … make standardization of

[6] The highest scoring system had an unweighted score of 748, a score of 914 with the preferred weighting, and 1266 with the original almost odds ratio scale. The preferred weights increased the base score by 1.21, the original weights by 1.69. The relative increases in scores for the lowest scoring system (341) were factors of 1.01 and 1.11 respectively. (Stone, 1908, p. 98, Table XXXVIII)

> yearly work possible, would show exactly the place, manner, and amount of development of any particular ability, and would give a rational basis for the estimation of the influence exerted by any method, material, or teacher. (Courtis, 1909a, p. 58)

While it is possible that other teachers of the time were interested in documenting and understanding the development of mathematics learning, Courtis appears to be the first to publish the trend across a school, triggered by the publication of the Stone report. As unfolds below, he had a significant impact on the teaching and assessment of arithmetic in a large number of schools as a result.

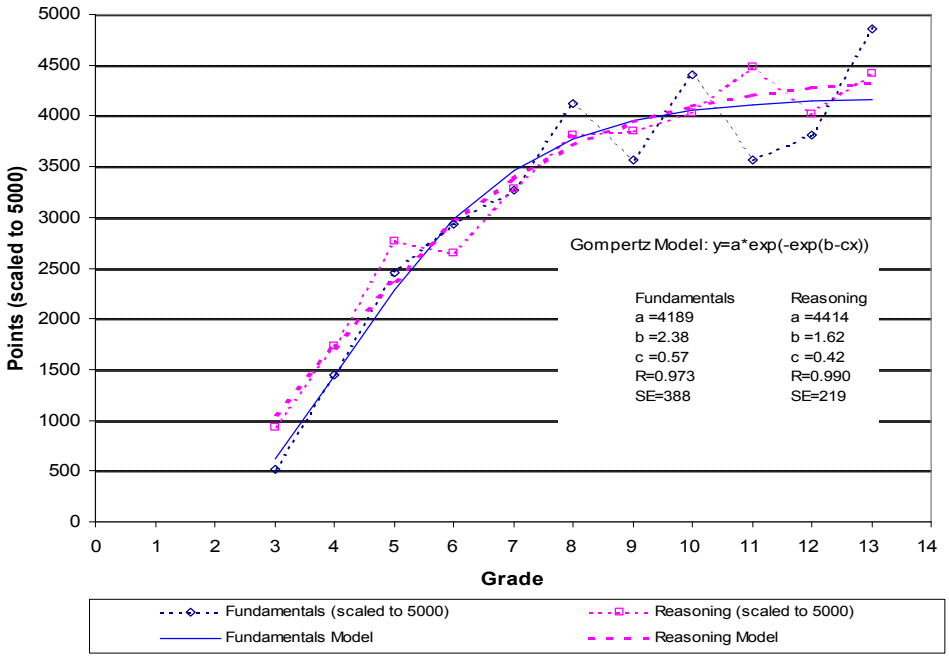
Courtis initially took a teacher and school perspective on the use of tests, and applied a new technology to understand learning at his school. Ultimately it led to a series of insights about the range of performance of students within classes and across classes, in what would appear to be the first published cross-sectional analysis of a school using a common quantitative assessment process. The test was clearly inappropriate for many of the younger students but some were able to complete the additions and solve some of the reasoning problems. Instead of using average scores (or median scores) for each grade, he presented the total aggregate score for 25 students per grade, requiring an adjustment to most cohorts to convert them to the 25-student standard.

He recorded the number of examples attempted, the number correct, the accumulated point value for the items attempted and the accumulated point value for correct steps taken, even if the final result was incorrect; a part credit scoring process. As an example, a question such as "divide 278542 by 679 is made up of 6 additions, 7 subtractions, 9 multiplications and 4 divisions" (Courtis, 1909a, p. 62) leading to a maximum possible score of 26 points. Courtis graphed the data to help the reader appreciate the apparent pattern of change with grade, showing an early growth of skill by grade view of arithmetic learning, along with variability across grades.

Courtis observed a pattern of alternating strengths in either reasoning or fundamentals by grade, one appearing to be stronger than the other at particular grades. He formed a view that arithmetic skills might be more fairly represented by a composite score, where the two aspects might balance each other out (in contrast to Stone who saw merit in keeping the aspects separate). Recognising that an unweighted comparison of fundamentals and reasoning grade point scores would be biased in favour of the more highly scored fundamentals (just under 5000 points per class) versus reasoning (just under 800 points per class) in combined raw scores, he rescaled both to a common scale of 5000 points (based on an assumption of a possible maximum fundamental score of 5000 for Grade 13); in what he called equating.

Figure 2.1 illustrates the result, redrawn from Courtis's (1909a) original data. He added hand drawn curves of visual best fit. Courtis showed (Figure 2.1), that grade variations in the relative skills of fundamental operation accuracy and success in reasoning problems, generally develop together. Although at any particular point one skill might appear stronger than the other, it may be an artefact of measurement error.

Figure 2.1 Points Scores for Correct Steps-Fundamentals v Reasoning



Note: redrawn from Table 11 of Courtis, 1909a, p. 71.

Smoothed trajectory curves hand drawn in original. Gompertz expression applied in example.

In lieu of Courtis's hand drawn curve, a computer-assisted curve-fitting process is applied to the data series for fundamentals and reasoning. The curve describes visually (and mathematically) the average trajectory of the development of these two skill sets, as reflected in Courtis's data, by grade. Using *CurveExpert* (Hyams, 2001) a range of curves can be tested for fit. The best fit is obtained with a Gompertz[7] model (Gompertz, 1825), satisfyingly appropriate as this model was ultimately chosen 17 years later, in 1926, by Courtis as the most likely model to describe controlled growth (Johanningmeier, 2004, p. 205). The curves establish a diminishing rate of growth in score points as Grade increases.

[7] The sigmoid model that best fits the Fundamentals data is a Gompertz model (SE= 388.3, R= 0.97). This fits slightly better than a logistical or MMF (Morgan-Mercer-Flodin) model. The sigmoid model that best fits the Reasoning data is also a Gompertz model (SE= 219.2, R=: 0. 99). This fits slightly better than a logistical or MMF model. See Appendix 5 for further information relating to the use of *CurveExpert* and the Gompertz model generally.

The idealised (i.e., fitted, smoothed) paths of development for both skills graphed in Figure 2.1 appear to follow, superficially at least, similar and almost coalesced trajectories for large segments. Whether this relationship would be sustained if the two aspects were equated using, say, procedures based on the Rasch model and a logit scale applied in lieu of rescaled points, cannot be determined. The means of raw scores for grades are assumed to correlate very highly with the mean of the Rasch model transformation of the individual scores, given the dependence on total scores in the Rasch model. The resultant increase in the spread of transformed scores is not likely to change the relationship of the two trajectories markedly.

Courtis's work was important in the development of quantitative educational assessment from the school perspective, as distinct from the system approaches of Rice and Stone. He appears to be the first to consider the time dimension of development using school grade. The publication of his analyses of data from the Liggett School for Girls, Detroit continued through 1910 and 1911 (Courtis, 1909a, 1909b, 1910, 1911a, 1911b, 1911c). Having used Stone's approach initially, Courtis moved to design his own instruments, which became popular very quickly.

By 1911 he had developed his own series of eight tests, the *Courtis Arithmetic Tests Series A*, and provided 30,000 sets throughout the US, England and Germany (Courtis, 1911c). He recognised the value of reference data being provided back to schools and aggregated 9000 individual scores across 14 grades from his tests to show patterns of typical development by grade (Courtis, 1911c). By 1913 with over 55,000 cases analysed for grade averages, he described the purpose of his tests as enabling the study of arithmetic abilities. He ultimately designated his initial test as Test 7 and designed simpler tests to lead up to this level of difficulty. From his analysis of the mistakes made by students he identified the "necessity for diagnostic tests of the simpler component abilities … and tests Nos. 1 to 5 were constructed" (Courtis, 1913, p.329).

By 1914 Courtis had broadened his view of useful data for observing the student and classroom. Under the slogan of "Measure the efficiency of the entire school, not the individual ability of the few" (Courtis, 1914, p. 380), he packaged a range of tests for English language development, covering handwriting quality, legibility and rate, composition generally, punctuation, spelling and syntax along with tests of memory. The handwriting tools were adapted from other authors (Thorndike, 1910 and Ayres, 1912) but other elements were of his own design.

In discussion of his approach to standard tests in English, Courtis anticipated a version of Vygotsky's Zone of Proximal Development (Vygotsky, 1978, p. 86). Based on his observations in English and Arithmetic, Courtis argued "it is not possible radically to change

the efficiency of present methods *until the actual work assigned to each pupil is based on his measured needs*" (1914, p. 392, italics in original). Courtis further explained that his tests "will furnish objective standards that will serve as goals for the guidance of teachers and pupils, and as a means of detecting the peculiar weaknesses of individuals." (1914, p. 392)

He concluded that the spread of achievement of reading was very similar to that of arithmetic and that he "expects to find that the same general causes operate to prevent success, that the same factors determine efficiency, and that the same changes in methods of teaching will prove effective." (Courtis, 1914, p 390)

He also anticipated some consequences of the Piagetian insights on stages.

> It has been a puzzling fact of teaching experience that ability to reason and ability to be exact in abstract work seldom go together. He is inclined to believe that there is a psychological principle at work, which, if known, would solve more riddles than one in educational procedure. Whatever the explanation, statistical proof of the fact is given here … Accuracy gradually decreases through the grammar grades [i.e. elementary/primary] and increases through the high-school grades at about the same rate. If this result is confirmed by future tests, there is an important lesson here. If inaccuracy in grades 7, 8, and 9, is due to some natural cause outside of arithmetic proper, to insist on accuracy or to spend much time in working for it may be not only wasteful, but harmful. (Courtis, 1909a, p. 73)

Having explored arithmetic, and anticipating his next phase into English language, Courtis speculated in 1909 that subjects taught over successive grades could be understood developmentally or longitudinally. He saw his tests as providing "a connective thread of growth in the fundamentals of the subject that will produce a unity that is sadly lacking in all present pedagogical effort." (Courtis, 1909b, p. 199) To understand student development over a broader time spectrum, Courtis recognised the need for observation and analysis that could connect across repeated tests and across the grades of the school. Single tests within grades, offered little if any insight to student learning development over time unless the tests could be connected to each other in some way.

In 1916 he reported 455,000 tests were sent out in one 12 month period (Courtis, 1916). His Teacher's Manual (Courtis, 1917) for the use of the practice arithmetic tests made clear the link he saw between testing and classroom practice. He offered advice on the efficient use of his tests to select who in each class should be involved. He targeted his support to the level of each student and encouraged teachers to adjust "the general method to … local conditions" (Courtis, 1917, p. 2). He summarised the steps in the use of his approach as

> a. Measure your class to determine the initial ability of its members.
> b. Eliminate from the drill class those who have (or reach) standard ability.
> c. Give to each of the other members drill upon those lessons where drill is needed.
> d. Permit each individual to practice in his own way and to grow at his own rate.

e. Give exactly the assistance needed to each child that fails.
f. Measure the efficiency of your teaching. (Courtis, 1917, p. 2)

His strategy for the use of measurement fits with the spirit of the arguments herein about how a teacher, having the best estimate of a student's learning status, might be expected to respond today, all-be-it that the range of supports should now be wider.

Courtis will be revisited briefly in Chapter 5 where further development of his concern about the relationship of learning growth with time is considered. At this stage the key insights from the early Courtis work are:

> teachers can be supported to use standard scientific processes to understand what is happening in individual student learning and within cohort learning;
>
> observing a range of aspects of learning developing together over grades enables their development with time to be understood; and
>
> many teaching interventions seem not to influence the rate of development yet students eventually improve.

Courtis's work could be labelled evidence based in today's terminology but he later became sceptical about tests. He withdrew his tests from the market in 1938 after twenty million copies had been sold because he "discovered they did not measure what they were supposed to measure" (Johanningmeier, 2004, p. 205). Courtis argued that repeated measures of the individual's "progress in terms of his own growth curve" were more useful than external norms and that growth was "cyclic in nature" (Johanningmeier, 2004, p. 206). His work provided an important impetus in encouraging teachers to observe the development of their students from a scientific perspective. (In a Frankensteinian escape, the tests took on a life of their own – independent of the intention of their creator.)

1910 Thorndike and the handwriting scale

Courtis's exploration of arithmetic coincided with an explosion in the range of assessment tools. One example was Thorndike's handwriting scale (Thorndike, 1910), which Courtis adopted into his English language assessment suite. A set of examples of handwriting was provided, each with a scale value that placed it on the scale at equally spaced intervals of quality (in Thorndike's view).

Thorndike (1910) published the scale in the *Teachers College Record* but was a little vague about the exact process of derivation. From pages 4 to 7 of the article it can be inferred that he followed the following steps: He selected a thousand examples of handwriting, many supplied by Rice, that were then rated into about 11 groups by 40 judges. As a result, each

example achieved an average score over all the judges, in the range 1 to 11. Thorndike selected examples close to the averages of 1, 2, 3, 4, etc. as his final examples and argued that they were about 1 unit apart in improving handwriting merit. To check the selected examples, he followed a second process of equally often noticed differences, where the judges compared the selected examples in paired comparisons. He explained "only if differences are not always noticed can we say that differences equally often noticed are equal." (Thorndike, 1910, p. 6) On this basis an example can be described by the percentage of judges who found a<b, a=b or a>b. Thorndike required his examples to show a separation of about 75:25[8] to justify a one-unit difference in the scale position. He also argued that 10 to 15 examples adequately spaced would be sufficient for a teacher to place an example of a student's handwriting at either of (or between) two scalar examples.

Thorndike considered the issue of where to place the lower and upper reference examples and argued that weaker and better examples than those on the scale of interest should be provided.

> The scale extends in actual samples by children from nearly the worst writing of fourth-grade children (quality 5) to nearly the best writing of eighth-grade children (quality 17). Quality 7 is nearly the worst writing of fifth-grade children.
>
> The scale includes a sample of a copy-book model which is rated by competent judges as of approximately quality 18, two samples of fourth-grade writing which are judged to be approximately of qualities 6 and 5, and a very bad writing, artificially produced, which is rated by competent judges as of approximately quality 4. The scale thus extends from a quality, better than which no pupil is expected to produce, down to a quality so bad as to be intolerable, and probably almost never found, in school practice in the grammar grades.
>
> If one had a finer scale, its use would give but slightly more accurate results, and would require more practice and more time. (Thorndike, 1910 p. 8)

The degree of fineness of the scale is considered again, later in the chapter, in the work of Hillegas. Thorndike argued the scale was necessary to be able to measure differences in the quality of handwriting. In a variation of the oft reported aphorism he claimed the

> …history of the judgments of the merit of handwritings supports the claim that if a number of facts are known to vary in the amount of any thing which can be thought of, they can be measured in respect to it. Otherwise, I may add, we would not know that they varied in it. Wherever we now properly use any comparative, we can by ingenuity learn to use defined points on a scale. (Thorndike, 1910, p. 69)

Thorndike saw the handwriting scale as a reference, something to which a teacher might need to refer in the assessing of the quality of any student work, and that through use, the scale

---

[8] The natural logarithm of 3 (75/25) is 1.09, that is approximating one logit, indicating that his scale (based on judges) has an approximate relationship of one unit to one logit.

would become internalised (that is referenced often enough to confirm an ongoing personal judgement calibration). Thorndike argued that the scale could be used as a mental standard, in the same way as an estimate of length might be made without using an actual ruler but by a tacit understanding of length units. He envisaged users of the scale having a stored impression of the quality examples.

This is the essence of the concept of teacher judgment of educational development considered in this thesis: applying a method that gives numerical substance to qualitative descriptions or categories. A framework is developed, used, internalised and the teacher's judgements becomes calibrated to the scale, with infrequent checking back to the original calibration examples.

1912 Thorndike's concept of scaling

Thorndike (1912), addressing the Harvard Teachers' Association, argued that educational science was able to apply the same process as led to physical scales to a wide range of educational developments. "Scales, graded standards, by which to report knowledge of German, ability to spell, skill in cooking, original power in mathematics, appreciation of music, or any educational fact you may think of, are now where the thermometer, spectroscope, and galvanometer were three hundred years ago - they do not exist." (Thorndike, 1912, p. 291) Using the scale for weight (in his example, in grams) he argued for four elements of an ideal scale:

> A series of perfectly definable facts; …
> Each amount is a different amount of the same kind of thing; …
> Differences between any two amounts are perfectly defined in terms of some unit of difference; …
> The zero point is absolute, it means 'just barely not any' of the thing in question. (Thorndike, 1912, p. 291)

He then described a range of educational development areas where "it is an easy task, theoretically, for educational science to take … vague, ambiguous statements of common-sense and refine them as physical science has in the past refined similar measures in the case of physical facts." (Thorndike, 1912, p. 292) Drawing on the concept of difficulty he explained that "in the case of spelling, we can define a point on the scale as the ability to spell words as hard as, but no harder than, 'a' and 'go', or 'wish' and 'touch,' and so on to 'millinery,' 'development,' or words of any difficulty we choose." (Thorndike, 1912, p. 291, commas as in original.)

Thorndike went on to explain the method of equally noticed differences, derived from Galton and Cattell and as used above in his handwriting scale. He used as his example the composition scale under development by Hillegas (of more, later) as an example of how

passages of writing of varying qualities could be rated by judges to create a scale, similar but more complex in its concept, than his handwriting scale.

Thorndike anticipated at least "two or three objections" (p.299) that teachers and administrators might have with a scaled approach:

> …the good old adjectives are enough for educational work
> …the common-sense judgment of a first-rate man without these units and scales is better than the action of the stupid man or incompetent man, with them.
> …the personal, spiritual work of education - the direct human influence that the pupil may get - is not in the domain of exact science. (Thorndike, 1912, p. 299)

He countered the first objection (existing adjectives good enough) with the assertion that for the kind of person making that objection, the use of the vague descriptors approach would suffice. As to the second objection, he acknowledged that a knowledgeable person without the scale might make better judgements than a "stupid man or incompetent man" with it but that it was the work of science "to get good work done by those of us who are rather mediocre". (Thorndike, 1912, p. 299) To the third objection he argued that the benefits of measurement and precision were not in conflict with more ethereal matters. "Mothers do not love their babies less who weigh them. We do not serve our country less faithfully because we take its census" (p. 299). While not exhaustive of the arguments of the 21st Century, his broad sweep covers some of the current concerns that measurement and judgment evoke.

1912 Hillegas: judging the quality of prose

The next application of Thorndike's scaled (holistic) judgment approach as applied in handwriting was that by Hillegas (1912) who, as a student of Thorndike, created a scale for the judgment of the quality of prose composition. The process is instructive in the labours taken to achieve this scale.

To start he acquired 7000 composition examples from "various sources and represent a definite attempt to obtain particularly the very poorest and the best work that is done in the schools" (Hillegas, 1912, p. 22). From these he selected 75 examples, supplementing the upper and lower ends of the scale with manufactured examples. The lower end samples were created by adults consciously trying to write poorly and the upper end from the youthful writing of Austen and the Brontes. Thus he started his calibration with 83 examples. The examples were typed with all characteristics retained (misspellings, punctuation etc.) to avoid the quality of the handwriting complicating the assessment, and then duplicated. In the first phase 100 judges were requested to rank the 83 compositions from worst to best and signify the order by numbers 1 to 83, or fewer if ties were required. Only 73 judges were able to follow the instructions correctly.

From the first responses Hillegas selected 23 compositions based on a process that selected the examples on the basis of steps in merit. He achieved this by selecting the poorest (about 95% of the judges agreed on this case), and the next two weakest examples. The balance of the examples were selected by finding the first case relative to the previously selected case that 75% of the judges had selected as better. Two gaps between the three weakest examples in the range were filled again with artificial samples. His aim was to create a scale that met Thorndike's ideal with a well-ordered set of examples spaced at exactly one Thorndike unit.

A revised set of 27 scripts was then sent to over 100 additional judges (teachers, authors, literary workers). In the second iteration the task was to create an ordered pile with the best at the bottom, the worst at the top, to be returned securely fastened. The first 75 to return were tabulated. Thorndike meanwhile had obtained 41 additional judgments from "individuals who were especially competent to judge merit in English writing", which were tabulated separately so their rankings could be used a "check on the others" (Hillegas, 1912, p. 40).

Additional judgments were also solicited from the general science community through an article in *Science* of June 1911 by Thorndike (Thorndike, 1911), which included only 8 of the original 27 cases for ranking, plus the latest two additions. The fate of these responses is unclear. Hillegas acknowledges 515 sets of responses overall, 202 of which were used in his analyses.

The result of all this labour was a set of 10 examples for use in the judgment of English compositions, the Hillegas Composition scale. It appears to be the first scale after the Thorndike Handwriting scale, to be offered as a tool to provide teachers a method to calibrate their judgment of composition merit. The items had values of 0, 183, 260, 369, 474, 585, 675, 772, 838, and 937. The scale units were 100 times the 'raw' unit that came out of applying the Thorndike scaling process. As described earlier this process assumed a unit of 1 (or 100 on the scale above) for a case where exactly 75% of judges rated a case as superior to a lesser case, based on Thorndike's use of the Median Deviation[9]. Thorndike (1916a, p. 228, Table 59) created tables to convert the difference in percentage of judges, working to two decimal points of precision, which one assumes Hillegas had used to look up the values. The implied precision alone may have been sufficient for many teachers to be sceptical about its utility.

---

[9] The Median Deviation is the median of the set of absolute deviations from the Median. It has a regular relationship to the Standard Deviation, with the constant dependent upon the type of distribution. For a normal distribution the SD is approximately 1.486*MD (now usually referred to as the MAD -Median Absolute Deviation). (http://en.wikipedia.org/wiki/Median_absolute_deviation)

The scale's purpose was to provide a ruler for a 'holistic' quality judgement. The actual characteristics of composition merit were not teased out; that is the composite elements of quality were not identified or made explicit. This led to some criticisms of the scale. Hillegas devoted his closing paragraphs of his 1912 article to defending his choice of multiple types of writing in one scale, requiring a holistic judgement. His defence of his approach was that

> people actually did do it. Of the four hundred and fifty people who have judged these samples not more than three have offered any objection on the score that they could not compare the samples. (Hillegas, 1912, p. 55)

Others were enthusiastic in their promotion of the Hillegas scale for system and classroom use. Abbott, of Teachers College, Columbia University, distributed copies of the Hillegas-Thorndike Scale at the 1916 conference of the National Council of Teachers of English (National Council of Teachers of English, 1917). He reported that through the use of the Hillegas scale the variations in the markings of freshman essays were greatly reduced, but that the markers felt the need of a scale to distinguish between form and content. It was reported that the scale had been used in Salt Lake City where fourth-grade pupils attained an average of 29 (the last digit had by now been deleted); fifth grade, 31; sixth grade, 38; seventh grade, 44; and eighth grade, 54. This was interpreted to show that steady progress in composition merit was being made. Cross (1917) advocated scales generally for the classroom teacher and believed the Hillegas scale was effective. "It would seem that there is too much room for individual opinion in judging here; but in the experiments I have made with the scale, having a number of persons read the same composition and then grade it by the Hillegas scale, the results were much more nearly uniform than I expected." (Cross, 1917, p. 188)

Thorndike (1913) considered the issue of errors of judgment using the scale, and speculated on the likely behaviour of teachers using the scale. He considered that initially errors would be large but that they would "diminish with practice in using such a scale and with improvements in the scale itself". With practice, he believed, errors would be "smaller than the errors now made by teachers in grading paragraph-writing for general merit" (Thorndike, 1913, p. 556). He argued that the reason for the errors being smaller was that a teacher, in grading a composition for general merit, used a subjective, personal scale of values that "cannot, on the average, be as correct as one due to the combined opinions of a hundred or more judges who are on the average as competent as he is." (Thorndike, 1913, p. 556) Hillegas's scale, he claimed, eliminated the errors due to the personal scale altogether and with enough practice with it would probably decrease the errors of comparison.

Thorndike (1913) and Trabue (1917) provided additional or alternative items for the Hillegas scale (creating the Thorndike Supplement and the Nassau County Supplement). In principle these samples were use in an attempt to maintain equivalent difficulties to the original scale

so that the scale was preserved. Trabue (1917) argued that the original scale, while helpful, had some deficiencies: artificiality of the lower examples, the brevity of the examples, a need for examples to be more similar to the type written in Nassau County, and, finally, some indicative data of standards by grade were desirable.

These early prose judgment scales, while controversial and imperfect, confirmed that it was possible to quantify classroom phenomena on the basis of teacher judgment applied to authentic student work, as long as teachers were provided with a reference frame or scale. Subsequent scales developed in the same period were more specific, that is targeted to very particular skills, for example, punctuation. There was also a trend towards conventional test schemes, those concerned with performance at the point of testing for selection or grouping purposes. This test score tradition is described by Engelhard (1991b, 1992b) as developing strength from the work of Wood, also a student of Thorndike. Wood was "a driving force behind the measurement movement of the 1920's that replaced essay examinations with multiple-choice items" (Engelhard 1991b, p. 146).

1913 Criticisms of scaled approaches from researchers of the period

While the scales had their advocates, they also had their critics (Johnson, 1913; Thomas, 1913; Learned, 1913; Neilson, 1913; Thurber, 1913; Holmes, 1913).

Johnson (1913) set three different groups (N=42; 16; and 5) the task of applying the Hillegas scale to eight composition examples. The range of variation from the lowest to the highest allocated values for any one item, averaged over all examples, was 3.7 Thorndike Units. The mean scale values for each item from each of the three groups, however, were close. The mean ratings for each example for each group correlated with the other two groups at between 0.98 and 0.99 (author calculation) indicating that while the within group variation was large, the orders of the average scores in each group were very consistent. Based on the upper and lower range values (the only exact data points reported) it is possible to observe that judges assigned values outside those of the examples in 65% of these cases. Judges used the scale as Hillegas and Thorndike expected and were not limited to the specific example values.

Learned (1913) reported an investigation where 50 papers were graded by 15 teachers. Initially the papers were graded using a percentage scale, then a month later with the Hillegas scale. The spread of values was reduced to 75% of the original range for all judges when the Hillegas scale was used, and to 56% of the original range for the 9 judges closest to the median. Thurber, in the same leaflet, (Thurber, 1913, p. 7) took a strong negative position, which on the face of it could have been a tongue in cheek argument for scales, though from the context this was unlikely to be so.

> The most baneful effect of the use of scales is that they inevitably make theme correcting more objective, and less subjective; the teacher's attention is at once focused upon the paper and not upon the boy who wrote it,—upon abstract qualities of writing, not upon personal qualities of the writer. The Hillegas Scale, as any number of better scales, used ideally, would make it possible for any English teacher in the country to correct and mark papers exactly as well as the teacher for whom those papers were written. Such a thing, on the face of it, is absurd. (Thurber, 1913, p. 7)

The Thurber comment highlights one of the tensions in the use of scales and judgement. The assessor is focussed on the merit of the student product. In principle, the same product should be judged by all teachers to be positioned at approximately the same place on the scale. Thurber's reason for the assessor to take into account "the personal qualities of the writer" in the judgement is not clear.

The Hillegas scale was still being used in educational research as late as 1940 but not as a classroom scoring tool or scale but as a method to identify text pieces of differing quality (Hinton, 1940). The process to develop the original scale was comprehensive and perhaps more complex than it needed to be. The Hillegas scale had some utility but it did not survive. It illustrates the principle that reference examples can be used to provide a score to a piece of writing and that values selected by assessors are not limited to those of the examples. Assessors used the examples as a scale and estimated values in between scaled examples. The principle of holistic marking of essays has been retained in modern US testing processing, although with less refined scales and, on occasions, less sophisticated markers than experienced classroom teachers (Farley, 2009).

1914-1916 Thorndike's scaling for reading

Thorndike (1914b, 1915, 1916b) applied his scaling approach to two elements of reading: reading words adequately to categorise them, and silent reading of passages of increasing complexity and then answering comprehension questions of varying complexity. Gray, another Thorndike student, developed at the same time a set of reading passages for reading aloud covered in the same article (Thorndike, 1914b). Scales were developed for these three aspects of reading development.

Thorndike assumed invariance of item difficulty over time and across locations for all items. He acknowledged however that there were local variations in difficulty. Words unfamiliar in one region may be commonplace in another, though he argued that these variations were usually exceptions and had only a small impact on his general approach. This general invariance of the difficulty structure of sets of items is a fundamental requirement if any scale of learning is to be feasible.

Thorndike's empirical approach was time intensive to develop and thus beyond the resources of classroom teachers. The product, however, was readily applicable in the classroom and by the 1915 version able to help assess the skill development of individual students. A benefit was that "the results will be readily comparable with thousands of others obtained with other classes by other supervisors, and will be at once understood by anybody who knows the scale—a most desirable feature" (Thorndike, 1914b, p. 14). His assessment approaches were eclectic, encompassing test items for students (words of established difficulty) but also holistic judgements (the Thorndike writing scale or the Hillegas scale). He was not averse to this use of observer judgement, believing he had processes to scale judgements as well as word difficulty.

1916 Trabue's completion test

Another of the scaled tests, the Trabue Completion Test was the subject of further investigation by Engelhard (1984) and is described briefly as it indicates how well scales of difficulty were being applied. The test itself, based on the Ebbinghaus-developed idea of filling missing gaps in text, was used as an indicator of the ability to think about words and language. Engelhard (1984) considered the data reported by Trabue informative about the scale structures developed by Thorndike, and an alternative approach to scaling developed by Thurstone.

Trabue's test began with sentences so simple that a large majority of the second-grade pupils were able to complete them correctly, and finished with sentences so difficult that only a small percentage of freshmen in college could complete them; that is, he had a concept of ordering the test by the difficulty of items. With an interest in measuring progress from year to year or from grade to grade, Trabue established empirically the difficulty of each sentence by trialling the incomplete sentences with thousands of public school children. From the results he developed four approximately equal scales, each scale consisting of ten sentences. He explained that by measuring ability at the beginning of a year with one scale and then at the beginning of the next year with an equivalent scale, it would be possible to determine the amount of progress made by a class or by a child during a year (Trabue, 1916, p. 88).

Trabue explored partial credit scoring as part of his analysis of the performance of the test. He tried initially six grades of quality (5-4-3-2-1-0) in the completion of a sentence. Trabue found that nothing was lost by "simplifying the scoring still further, giving two points credit for each perfectly completed sentence, one point for each sentence completed with only a slight imperfection, and zero for any sentence omitted or imperfectly completed". This "had benefits in efficiency of marking with no loss of information for scoring students" (Trabue, 1916, p. 87). He viewed items as being linearly scaled in difficulty, with the point where

students were unable to add an appropriate word being the measure of current student ability, implicitly seeing students and items on the same scale and also implying item invariance (Trabue, 1916).

Item Difficulty and the key link to educational measurement

Were the items spaced along a continuum of difficulty in a fashion comparable to that might be established today?

Engelhard (1984) acknowledges that the problems and issues in psychometrics have not changed much since early 1900 and were well considered by Thorndike, and later Thurstone. The conditions necessary for objective measurement as described by Rasch model advocate, Wright (1968, cited in Wright, 1977) include: the calibration of measuring instruments must be independent of those objects used for calibration, and the measurement of objects must be independent of the instruments used. These conditions require the objects/items used in calibration of scales to be invariant in relative difficulty across samples or occasions.

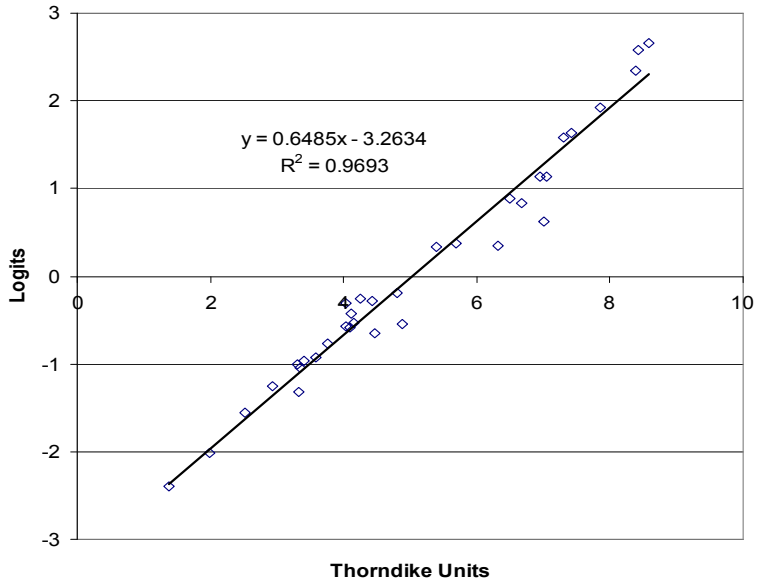
Thurstone (1925, p.433) commented on the inadequate discussion of the "assumptions and the logic of … scale constructions" and proposed a new approach to scale construction. He illustrated his process with a scale he developed for the Binet test questions, re-analysing Burt's data from 3000 London school children. His analysis, also based on standard deviations, established a scale of order of items for the Binet test questions, with an origin at the mean of Binet Test intelligence for $3^1/_2$ year olds. The result was the first published item map. It illustrated graphically some characteristics of the Binet test unappreciated up to that time. These characteristics included confirming a general spread of difficulty along the standard deviation based scale, but with major gaps (scale segments with no items) at the upper end of the scale. The scale highlighted a strong bunching up of items at about 2.5 units on the scale. The elegant ruler like presentation (Thurstone, 1925, p. 449) embodied the concepts needed to understand how items could be used as markers of learning development. The detachment of the scale from age cohorts to an absolute scale complements the exemplar scales of Thorndike. The graphic and the scaling process depend on the adequacy of the standard deviation unit as an appropriate unit.

Thurstone (1928) required that "the scale value of an item should be the same no matter which age group is used" (p. 119) and believed that Thorndike's approach was dependent upon the samples used, and thus not sample-distribution free. Engelhard reports that the Thorndike and Thurstone approaches yielded "essentially identical values when applied within one group" (Engelhard, 1984, p. 31). In multiple groups, Thurstone adjusted differences in the means of different groups. Thorndike assumed that the standard deviations in each of the ability distributions were equal (Engelhard, 1982; Holzinger, 1928). Using

Trabue's Completion Test data, Engelhard established that Rasch, Thurstone and Thorndike scaling processes produced linear scales of item difficulty that, within and across methods, were approximately invariant, and that each process had a linear relationship with the other two.

Plotting the data reported by Trabue and re-analysed by Engelhard, the relationship of Thorndike Units and logits is illustrated in Figure 2.2.

Figure 2.2 Trabue's Completion Test from Engelhard (1982)



Note: Data points are item difficulties for Trabue's Completion Test in Thorndike Units and as logits, based on Engelhard's (1982) conversion.

For Trabue's completion test data there is a strong linear relationship of Thorndike Units to logits. The scale analyses confirm that both Thorndike's and Thurstone's scale approaches are transformable to a logit scale. This confirmation implies that there is a consistency in their scaling approaches developed in the first quarter of the 20th century with those based on the more modern Rasch model. These early insights into scale concepts have yet to be fully realised in approaches to classroom assessment. However the progress map initiatives described in the subsequent chapters draw on this vision of scales, as do, less directly, the concepts of levels.

1916 A 'level' approach for composition

Hudelson (1916, p. 595) advocated the use of scales and urged that "we must start somewhere, and rather than go through all that has been done to work out established scales, we can use such standards as the Hillegas or the Harvard-Newton to fix our units of

measurement, much as the zero and boiling-points are established on a new thermometer by measuring it with an authentic one." (Hudelson, 1916, p. 595) He offers that

> for composition, by establishing one or two points we can, from them, derive the other degrees. These need not be fixed upon a percentage basis; in fact, my belief is that a standard should be made to run from 10 to 120 in point or degree of difficulty; and the ideal minimum for the first grade would then be 10, for the second 20, for the third 30, and so on, up to 120 for the twelfth year. With this as a basis, our problem would then become one of choosing typical models for each year. (Hudelson, 1916, p. 595)

This is an encapsulation of a vision that was taken up (or re-invented) much later as part of the curriculum and outcome descriptions encapsulated in the English and Australian curricula in levels models in the 1980s and 90s. In these models the assessment vision involved processes to help the teacher assess either where students are in their development (where they are located on the scales) or where an artefact (item) produced by the student is located. The Hudelson scale vision assumed equal linear increments of 10 units per grade. Assuming continuous improvement this would average about 1 unit per school month. The concept of an extended scale connecting the elements of the curriculum and scaling their difficulty, while not particular to Hudelson, anticipates later developments, particularly Wright's 'Academic Achievement Units', where, "say, 0 = entry into 1st Grade and 1000 = admission to College." (Wright, 2001, p. 784)

1950s - A new way forward.

In the period from the 1920s through to the 1950s the use of group and individual testing approaches increased. A range of journals was founded (*Psychometrika*, 1935; *Educational and Psychological Measurement*, 1941; *British Journal of Statistical Psychology*, 1947) (Du Bois, 1970) and a rich exchange of approaches to the development and analysis of test processes developed.

Somewhat isolated from this mainstream development, Georg Rasch, working mainly as a statistical consultant, was engaged to assist in the development of an intelligence test for the Danish military (Andersen & Olsen, 2000). This initial encounter with test and item difficulty led into a project on slow readers where children had been tested and remedially supported in their school years, and re-tested as adults in 1951. For various reasons (different reading tests, World War II) "it was not possible to evaluate the slow readers by standardisation as was the usual method of the time." (Andersen & Olsen, 2000, p. 10). Rasch needed to develop a method where an individual could be measured independently of which reading test had been originally used and in a way that could be connected to the 1951 test.

> The method was as follows: two of the tests which had been used to test the slow readers were given to a sample of schoolchildren in January 1952. Rasch graphically compared the number of misreadings in the two tests by plotting the number of misreadings in test 1 against the number of misreadings in test 2 for all persons … The graphical analysis showed that apart from random variations, the number of misreadings in the two tests were proportional for all persons. Furthermore this relationship held no matter which pair of reading tests he considered. (Andersen & Olsen, 2000, p. 10)

To account for the random variation Rasch developed a Poisson model. He was able to develop a model with two parameters, one for the subject (or person) and one for the item. Rasch had established that performance of students on a test could be related to the difficulty of the test and that it was "possible to deduce a distribution that depends only on the item parameters, but not on the person parameters" (Andersen & Olsen, 2000, p. 13). Through this process Rasch developed an approach to "bridge building", the process of placing persons on a common scale using different test instruments.

Subsequent consideration of similar problems led to the development of the Dichotomous Rasch model though the timescale for this development is not clear (Andersen & Olsen, 2000). The model, sometimes called the one-parameter logistic model (Allen & Yen, 1979), has become the basis for a family of models that maintain the independence of person from item.

This simple approach, when data fit the model, has opened up a range of possibilities that support the assertion that the placement of students on a (latent) scale can be estimated via a wide range of instruments or processes. The development of scales using the Rasch model makes feasible the use of teacher judgement as one of the 'instruments', as already described and advocated by others (Forster & Masters, 2004; Griffin, 2004). Engelhard (1984) established that the natural unit of the scale, the logit, already applies in the earlier scales developed in the period 1910 to 1930.

Summary

The early explorations of Fisher, Rice, Stone, Thorndike, Courtis, Hillegas, Trabue and Thurstone, as described above, set in train a technology and a concept of how learning in the classroom might be monitored in a consistent fashion. The approaches led to tools that teachers themselves could use in the classroom to monitor student development. Some of the developers were interested in the change of learning performance with grade and presented data that helped teachers and school administrators see assessment as an approach to observing and confirming individual student development over time. In many cases these data assumed latent scales of development, calibrated with equal interval units that bore (it can now be established) an approximate but direct relationship to the currently widely used

Rasch logit scale. For the useful observation of development over time, the learning unit must be of a consistent amount. That early developments in scaling bore a strong relationship to the now accepted log odds scale provides confidence that these early researchers were creating measurement scales with meaningful units.

This initial direct use of scaled techniques to assist holistic judgments by teachers and establishing estimates of scale positions of students on other developmental scales, seems to have lasted in this form until the late 1920s but was not well embedded in classroom practice. Assessment moved from a scaling tradition (Thorndike) to a test score tradition (Wood) in this period (Engelhard, 1992b). The production of an increasing range of paper and pencil tests was combined with the tendency to narrow quantification from measurement with extended rulers of learning, to become focused on pass-fail tests that established who crossed critical boundaries. This was achieved without optimising the meaning of the value of the underlying latent scale. The concept of a measurement scale was disregarded altogether by many teachers who instead perpetuated alpha grades, adjectives and/or percentages. With the development of approaches to assessment, particularly those initiated by Rasch in the 1950s and 60s, concepts of educational measurement were redeveloped. These led to the possibilities of extended scales of development and, in some implementations, the use of teacher judgment as the process for establishing the position of students on scales of development.

Although limited by manual data processing, and a reliance on paper records in the classroom, the early scale developers provided a process to support data informed pedagogy. The early beginnings of the conceptual steps in educational measurement, while concerned often with monitoring teachers more so than learning, have set a path that can be redirected back to helping teachers assess consistently and to record assessments in a systematic way. That the potential was not fully developed can be interpreted, by some, as a failure of the processes. For others, and in terms of this thesis particularly, the evidence is that appropriate thinking and concepts have existed for 100 years. Scaling and mapping learning, critical still in test technologies, has the potential to be applied to help teachers better understand and thus better observe and assess the learning of their students.

To summarise the perspective of the 1920s, McCall in his *How to Measure in Education* (McCall, 1922) argues that there are many reasons to see teachers' judgements (of the period) as inadequate or inaccurate in classifying students; this role of classifying now no longer a prime requirement of the classroom. However he comes to the view that teachers' judgements have importance.

> Teachers' marks are important because they are now and will continue for some time to be the most universal method of rating pupils. In fact, they may continue forever to

> be the criterion for classification [in the modern context read 'assessment'], because teachers will soon be familiar with the simple mysteries of scientific measurement. They will themselves use tests with the same ease and fluency that they now use textbooks. More and more they will base their judgments upon objective rather than subjective measurement. When this time arrives teachers' marks will be not only as accurate as objective measurement, but they will be objective measurement plus something else. (McCall, 1922, p. 59)

To achieve this accuracy outcome requires support and encouragement for teachers, with the right frameworks and assessment tools. Curricula in levels combined with processes for assessing progress through each level, provide one model for doing this. The next chapter considers the more recent history of the development of levelled curricula, particularly in Australia in the 1980s and 90s that led to the teacher judgement data analysed in Chapters 7 and 8. A curriculum described in levels bears a conceptual relationship and a direct historical link (based on Hudelson) to the scale developments. A key attribute of early scales is the process to position individual and group summaries of learning status at scale values between the 'prime' scale markers. Initially, this attribute did not carry over to the Australian levels, in South Australia at least. A process to estimate progress from one level to the next is required to maximise the value of a level scale to teachers.

Chapter 3: Levelled curricula, learning progress and skills tests

> In studying the child the teacher tries to learn his innermost thoughts so that she may be able to render her guidance intelligible to him. As she learns to understand him she begins to sympathize with him, and in return she secures his love; once his love is secured, he will follow her to the end of the earth, and the examinations will take care of themselves. Thus the weight of oppression becomes removed from the child; he becomes free and happy in his freedom, and the school is converted into the loveliest of homes.
>
> Rice, 1893, p. 97.

The previous chapter established that assessment schemes using scales and examples of known scale quality to estimate the learning status of students had their roots in the work of mainly American educationalists in the early 20th century. This chapter addresses the development of the Australian curriculum and assessments structures in the 1980s and 1990s, particularly as they impacted South Australia.

One distinctive element of the assessment process for these curricula was the requirement for teachers to judge how students were progressing. They were required to make these judgements using a curriculum framework structured in levels. This structure was a change from most previous curriculum descriptions, which described content to be covered in specific grades or Year levels. The level concept, a description of skills and other attributes classified into levels of increasing complexity, was key to the creation of the teacher judgement data, summarised in Chapter 7. In the mid 1990s South Australia also introduced statewide testing in primary schools. Data from the tests of the period feature in Chapter 6, and a comparison of teacher judgements using levels with the test assessments is made in Chapter 8.

The chapter outlines a brief history of the development of the levels structure for the proposed but only partly implemented Australian national curriculum of the mid 1990s and the strong link in that curriculum approach to teacher judgement as one basis for assessing student progress. South Australia contributed to the national developments while establishing its own level structure in the form of attainment levels, prior to the ultimate adoption of the national level structure.

The development of 'Profiles' and 'Levels' for Australia

Documents describing a level approach to curriculum were developed in the period from 1988 to 1993 under the auspices of the Australian Education Council (AEC) based on

recommendations from the Directors General Conference and the Australasian Cooperative Assessment Program (ACAP) (Lokan, 1997).

Progress for a student through the curriculum was described as eight levels of increasingly sophisticated skills and knowledge within eight learning areas. For each learning area there were two documents produced; a statement that provided "a framework for curriculum development" and a profile that "described the progression of learning typically achieved by students" (Curriculum Corporation, 1994c, p. 1)[10].

Aspects of the Statements and Profiles for Australian Schools (SPFAS) documents development process have been recorded by Lokan (1997), Marsh (1994), Piper (1997) and more informally by Jenkin (1996). Jenkin (1996) was well positioned to observe the developments in his role as executive officer to some of the committees involved in the developments nationally and in South Australia. Boomer, as Chair of ACAP and South Australian Associate Director General - Curriculum and formerly Director of the national Curriculum Development Centre played a key role in the initial development of the statements and the concept of teacher judgement assessment as an alternative to system wide and national testing (Jenkin, 1996). Boomer helped link local SA initiatives in attainment levels to the national approach.

Precursor approach- Attainment Levels in South Australia

During 1990 South Australia had been considering its own approach to a levelled curriculum in the form of attainment levels. Boomer commissioned the South Australian Curriculum Directorate to develop attainment levels (Education Department of South Australia, 1992) with the endorsement of the then SA Director General, Boston[11] (Jenkin, 1996). The brief for developers required 6 levels from Reception (R) to Year 10. Levels attained were to be standards referenced. Reports for parents based on these levels were to be provided along with a school and system perspective by curriculum area (Stehn, 1997).

Masters (1999) reports that Griffin's Literacy Profiles (Griffin, 1990) were "influential in shaping later initiatives to develop the South Australian Levels of Attainment and more

---

[10] For convenience these two aspects for each learning area are described generally as Statements and Profiles for Australian Schools (SPFAS) throughout the text, as introduced in Chapter 1.

[11] Boston subsequently moved to NSW as Director General. As chair of the AEC Curriculum and Assessment Committee (CURASS) he brought the national SPFAS document development program to its completion (Marsh, 1994).

significantly the national profiles" (p. 289). Both Masters and Griffin were influenced in their views on curriculum, learning and assessment by their understanding of the Rasch model (Masters, 2005; Griffin, 1998), underlining the conceptual impact of the Rasch model on aspects of the design of the South Australian attainment levels and ultimately the national profiles.

The implementation of attainment levels in 1992 meant that SA teachers were already exposed to some developmental concepts in curriculum design that would flow, in general terms, into the levels of the national profiles (Jenkin, 1996). The spin-off from the attainment levels to classroom assessment was less well developed. However, prior exposure to a similar concept meant that the lead-time into the eventual adoption of the profiles and levels approach in South Australia had a benefit of three years more exposure than applied for some other parts of Australia (Marsh, 1994).

The parallels in the two approaches are reinforced in a pamphlet of the SA Curriculum Division (Education Department of South Australia, 1993). The similarities justified the continuing "familiarisation programs for the attainment levels" in 1993 "when we are moving towards the adoption of national profiles". The similarities were that both were standards referenced, both valued teacher judgement, both provided a tool for teachers to describe student achievement and both were a description of the progression of learning typically undertaken by students in each learning area. Many SA teachers had a benefit in developing their understanding of national profiles from their experience with attainment levels. Industrial action relating to teacher work load, described briefly later in the chapter (Stehn, 1997), influenced which teachers were able to maximise their use of attainment levels. At some sites attainment levels and then SPFAS were effectively banned.

Teacher judgement in Attainment Levels and SPFAS

Boomer believed in a system of assessment and reporting that supported the use of teachers' judgments and that valued their tacit understandings of their students. He had in mind, according to Jenkin,

> a system that would help teachers make more reliable judgments about students' achievements … without undermining their professional credibility and integrity. Accordingly the profiles framework was seen as needing to accommodate the way teachers and students actually worked together in schools and to be sufficiently broad not to impose a construct that limited the classroom options. (Jenkin, 1996, no page reference)

ACAP with Boomer as chair, proposed standards referenced frameworks based on the work of Sadler (1987). While generally similar to criterion referenced reporting, standards referenced frameworks were seen as drawing "upon the professional ability of competent

teachers to make sound qualitative judgments of the kind they make constantly in teaching" (Sadler, 1987, p. 193). This capacity for qualitative judgment was seen as being able to be "refined to the point where it could be used, directly, in the classification of student achievement into grade levels" (Sadler, 1986, quoted in Sadler, 1987, p. 193). Implicit in the Sadler concept was a holistic judgement of the quality of student leaning, not based on tests or mechanically accumulated ticks of outcomes met.

The adequacy of the level descriptions

The correct linking of outcomes to levels was required for the SPFAS to be effective. The Australian Council for Educational Research conducted calibration studies in 1992 and 1993. The studies considered the separation of levels, the equivalence of outcomes within a strand within a level and the ability of teachers to understand the assessment process (Lokan & Wu, 1997). The studies showed consistency in the pattern of the upper and lower thresholds[12] of outcomes by profile level and an even growth from one level to the next.

Masters, while acknowledging that the national curriculum was "developed hurriedly, and was best viewed as drafts of frameworks for curriculum, assessment and reporting" (1999, p. 280), suggested that the materials provided opportunities to address a number of empirical issues. In particular he asked, "Was the sequencing of outcomes along each strand, based on the experience of curriculum designers, consistent with the empirical ordering of assessment tasks designed to address those outcomes?" He saw an iterative process linked to parallel collections of test data as "useful for revising and refining the outcome frameworks" (Masters, 1999, p. 280). This never materialised.

The development of the statements and profiles was completed in June 1993 at a time of change of governing party in many state governments, which ultimately resulted in uneven implementation across the states. However, the initial development of the national (and Australasian) approach to a levelled curriculum was originally a cooperative initiative of the Directors General of each of the education systems of Australia, through their senior curriculum leaders, rather than an imposition from any outside entities.

Implementation in South Australia

The July 1993 AEC meeting agreed that any future publication of material would be the prerogative of each state and territory. This resolution was meant to imply neither

[12] Thresholds are at the point where students on average, when assessed for a specific outcome in a strand, move from 'hardly ever' to 'sometimes' category in that strand. The upper threshold is the point where students move from 'sometimes' to 'almost always'.

endorsement nor rejection, by the states and territories. South Australia proceeded to purchase copies of all documents for its schools. The professional development program for South Australian teachers had continued, flowing easily from the attainment levels approach to the SPFAS approach, although the transition was not without tensions. Marsh (1994) records that South Australia "made considerable advances with an implementation program, largely due to the three years' experience gained from developing and trialling state attainment levels" (p. 169).

Stehn (1997) comments, "perhaps the most difficult years in the history of the South Australian curriculum review and reform initiatives were 1992 and 1993. These were the years of metamorphosis … from attainment levels to nationally developed Curriculum Statements and Profiles" (Stehn, 1997, p. 173). The teachers' union had difficulty with perceived economic rationalist motives to the reform and as well the additional demands on teachers' time and a period of industrial unrest continued. Some schools struck deals to implement the profiles with staff, who were in many cases intellectually and pedagogically sympathetic with the general approach, seeing it as a re-packaging of current practice (Stehn, 1997). The union banned involvement in implementing either the attainment levels or the SPFAS.

Appreciating the complexity of the industrial situation it is not surprising that most SA departmental effort went into the explication of the impact on curriculum planning and attempting to ameliorate the perception of increased workload. As a result there was very little interest in exploring the broader student assessment possibilities including methods of recording a student's level status with greater refinement (author direct experience, 1994 to 1996).

Progress indications within a level

The intention to collect data on levels achieved by students in South Australia had been signalled early (Education Department of South Australia, 1993). However, progress within levels does not appear to have been discussed in operational detail by the South Australian implementation planners. This was revealed, by implication, in the arrangements that were eventually put in place for the collection of level data from teachers.

Progress within levels was an issue both for system-wide data collection and for teachers themselves. The time between attaining the criteria for one level and attaining the criteria for the next level was about two years for any student. How part progress towards the next level should be described, or if it was even desirable to do so, was left unspecified. Teachers, however, were adamant that some form of part progress record would be needed or else any data collections would misrepresent the learning that was occurring (Private conversations by

teachers with the author and negotiations with the teachers union in which the author participated, 1997).

The Victorians Curriculum Standards Framework (CSF), essentially equivalent to the SPFAS, addressed progress within a level by dividing the distance into three zones; 'beginning', 'consolidated' and 'established' (Department of Education, Employment and Training, Victoria, 1996). The South Australian Curriculum Division of the time was not enthusiastic about clarifying the detail of what the likely data generated by teachers should look like and what value there would be in collecting it (Author observation). This was partly because of the sensitivity of the issues that were related to getting data from teachers, in the context of the generally alleged economic rationalist approach and ongoing workload issues. Stehn (1997) reports that a number of related projects were underway: "Ten resource papers on issues such as programming, reporting and using student achievement information have been published" (Stehn, 1997, p 185) but in none of these was the issue of within level progress adequately considered.

The matter was also raised in ongoing evaluations of project implementation conducted by ACER in 1995 and 1996 for the SA Department (Frigo, 1997). Frigo reports that concerns identified included "the broadness of the levels, the slow movement of students through levels that don't account for progress made, consistency of levelling by teachers" (Frigo, 1997, p. 20). There was concern about the use of numbers as opposed to descriptive reporting. There were concerns about the self-esteem of older students assessed as being at lower levels than usual for a given Year level. It was alleged by one teacher that "there is no room for 'distance travelled' for the slower child, who is not ready for formal learning at 5, but may have made huge progress for his/her ability" (Frigo, 1997, p. 20). This specific criticism related to students below level 1 but the inability to record distance travelled applied at all levels. The design was intending to articulate distance travelled as one of its fundamental elements but the final product had not clarified how this might be done.

Even though some issues about progress within levels were flagged in the Frigo evaluation these did not flow into a consideration of what options there were to address them.

Data collection in South Australia

There had been a clearly stated intention to collect data from schools by 1995 (Education Department of South Australia, 1993) but this collection was delayed until 1997. In early 1997 officers of the Department for Education and Children's Services (DECS) began negotiations with a representative group of the South Australian Institute of Teachers (SAIT) about the parameters for a data collection. SAIT negotiators were clear what conditions were

to be met. Based on author memory (no documentation is available to the author) the conditions were essentially:

no teachers were to be identified at any stage,

no schools were to be identified in reports,

maximum teacher involvement with minimal time demands per teacher,

and the 'sticking point', a method to indicate progress within levels.

An existing protocol, the *Code of Conduct for Using Student Achievement Information* (1995), ensured that no students would be identified. It is important to establish at this point that these identifications related to the publication of any data that would identify an individual student or could infer identification from summary data. As early as 1993 the need to include the student identification code was made clear (Education Department of South Australia, 1993) as the link to other desired special population identifiers (gender, age, postcode, aboriginality, non-English speaking background, SES status, disability code, and Year level). Using student identification codes was an innovation of the late 1980s in the Statistics Unit, under the author's management, allowing an increase in the range of data summaries from one data collection. By 1997 it was commonplace for most statistical collections from SA government schools to be automated, using school based student records organised by student identification codes and transmitted to central office. The collection model for levels data finally adopted for 1997, and repeated in 1998, required student identification codes to enable student characteristics to be attached (age, gender, any particular sub-population identifiers, Year level). The agreed protocols ensured that these would not be published. Teachers were never considered for identification.

Negotiations for the collection broke down over the indicator of progress within levels. It was at this point the author and a colleague (Ian Probyn) were invited, as representatives of the Quality Assurance Unit of DECS, to assist in the development of collection approaches. It was unfortunate that previous attempts to discuss the approaches that might be adopted for progress within levels had not been taken up by the Curriculum Division, as it meant there were no trial approaches that had been field-tested available for consideration. The options rather hastily offered were the Victorian Model (3 Zones), a 4-zone model, a 5-zone model, and a 10-zone model. The initial meeting agreed, with surprisingly strong support from the union negotiators, to further consider a 10-zone model, once a collection process could be explained. The author and Probyn (who did the detailed development) returned about 2 weeks later with a prototype collection process.

The prototype software automatically selected the sample of students to be reported by each teacher. It had been agreed, previously, that all teachers would be involved, with a requirement of 5 students only on average to be reported per teacher. The software developed by Probyn, randomly selected one of four learning areas for each teacher. It then randomly selected 5 students for each teacher through interaction with the school's computerised student record system. For each student the strands of the selected learning area were shown on the screen. For each strand, the teacher was requested to identify the level most recently achieved and, by clicking on a continuous bar (of nine elements undifferentiated from the teacher's perspective), the progress towards achieving the criteria for the next strand indicated. The elegance of the solution to indicating progress was that it could be rescaled to any chosen divisions (2,4,5,10) of progress and did not require the teacher to consider a decimalisation of the scale, even though this was the result of a progress judgment.

The union representatives were sufficiently happy with the prototype, partly because of the automation and thus a simple response process, and partly because of the opportunity to indicate progress. They agreed that they would support a data collection of this basis. In this way the first profiles data collection was agreed to, and then carried out in 4th term, 1997. Other departmental officers who had been negotiating for some time were bemused at the ease with which the matter had been resolved. It is a matter for speculation only as to whether a more elegant process might have evolved on the issue of progress indication if more developmental work had been done over the previous three years.

The collections were conducted in Term 4 in 1997 and Term 3, 1998. The process allowed for four learning areas in each collection. After two years of data, a learning profile of the system in South Australia had been developed. A series of brochures, describing the data were provided to teachers, for each of the learning areas. These provided graphical descriptions of the trends in student progress over the year levels 1 to 8. In almost all strands in all learning areas, the general picture was the same. The median students in each Year level were on a straight line of constant gradient (about 0.4 to 0.5 of a profile level) by Year level (similar to that shown in Figure 7.2).

The distributions of student profile levels per Year level show increasing spread as Year level increases. The development of this overview of student development within strands, within learning areas, was based on teacher observation data alone. As far as the author is aware no similar data set, collected by individual student for a system-wide sample, had been

developed elsewhere[13] at that time. Rowe and Hill (1996) provide a very similar view but for a smaller sample of schools.

The general result was presented to the Australian Association for Research in Education conferences in 1998 and 1999 (Rothman, 1998, 1999). The collections were not continued following the Frigo Report of November 1998 (Frigo, 1998). In late 1998 and early 1999 the Curriculum Policy Directorate in DETE SA developed a draft writer's brief for a revised curriculum framework (Hornibrook & Wallace, 2001). Improvement of the Statements and Profiles documents was a requirement, responding to the Frigo Report (Frigo, 1998) and Withers Report of January 1999 (Withers, 1999). Initially in-house adjustment to the Statements and Profiles documents was anticipated. Following a change of Minister and Chief Executive, an outsourced revision was requested for the development of a suite of new documents and curriculum structures.

> There is no paper record available … that sources or dates the change in focus and process. However, data from interviews indicates that the change was as a result of discussions with the Chief Executive.
>
> According to data generated through the interviews, the appointment of a new Chief Executive to the department provided a different way of doing things. It was reported to the evaluation that he expected collective and connected action and expected that more minds and expertise would be brought to the task. The task was to be done in a more connected way and be done more quickly. Planning for the task was therefore mindful of the necessity to destabilise traditional working patterns and to make new connections. (Hornibrook & Wallace, 2001, p. 10)

SA moved into the next phase of curriculum reform, with the development of the South Australian Curriculum and Assessment (SACSA) Framework commencing in 1999. Eight levels were reduced to five, for the same development span, with no consideration of progress within a level (South Australian Curriculum, Standards and Accountability Framework, 2000).

Confirmation of the value of profiles – application in studies and student assessment

Meanwhile profile levels had already begun to be utilised in research. They were used in the 1996 National School English Literacy Survey, in a similar fashion to the use of the Victorian

[13] In principle the CSF 1, CSF 11 and VELS series of teacher assessment data (Victorian Auditor-General, 2009) provides a similar view. As far as can be ascertained the collection of this data in 1997 was by schools reporting means per Year level per strand rather than individual student data. Since 1998 teacher assessments have been collected electronically from schools (Department of Education Victoria, 1999).

Curriculum Standards Framework (Rowe & Hill, 1996). Masters & Forster (1997) used the level structure to 'map' the literacy skills of Year 3 and Year 5 students across the nation. The management committee for the project proclaimed the value of the approach in

> documenting the varied Levels of student achievement in those aspects of literacy which constitute the framework of the English curriculum profile: Reading, Writing, Speaking, Listening and Viewing. This is in contrast to the more limited scope of earlier national surveys of literacy achievement which were developed to gather data about the percentage of students unable to satisfy minimal levels of competence in reading comprehension. (Masters & Forster, 1997, p. iv)

The management committee also praised the value of extensive teacher judgement in the survey process.

The original profiles concept (AEC, Conference of Directors General, ACAP, CURASS) included a mechanism to assess students (and as a consequence classes and schools), directly by their teachers, on the basis of how students were making progress through the levels. Keeves (Keeves & Marjoribanks, 1999, p. 129) reports that the sequences of instruction underlying the SPFAS "have limited meaning unless there are underlying scales that would permit the assessment of student learning in the form that Masters (1982) had envisaged". Masters indicates "from an educational measurement perspective, this initiative to specify intended learning outcomes, to organise these outcomes into strands, and to describe eight levels of progress along each strand, meant that frameworks were beginning to emerge which could be used to guide test development and against which students' test performance might be reported" (Masters, 1999 in Keeves & Marjoribanks, 1999, p.289).

Embodied in Master's view of the benefits of the SPFAS was the value of the materials, not only as an explicit statement of desired outcomes from the curriculum but as a developmental ruler against which student progress could be charted. Test items and assessment tasks, while indicative of the skills achieved were of interest "only to the extent that they are useful vehicles for estimating the location of students on the variable of interest" (Masters, 1999, p. 285).

Keeves reports that the SPFAS had characteristics that "warrant the claims made of innovation, development and marked advance in a world context" (Keeves & Marjoribanks, 1999, p. 114). Features included "scales of learning and the benchmarks as levels on the scales facilitate not only teaching and instruction but also the assessment and reporting of student learning and development over time". (Keeves & Marjoribanks, 1999, p. 114). Not everyone could see the benefits of the SPFAS.

Criticisms of a level approach

While a body of pragmatic and somewhat 'cutting-edge' curriculum and assessment concepts underpinned the profiles and levels approach (Masters, 1990, 1999; Sadler, 1987; Griffin, 1990), the concept of profiles had critics from educationalists. (Reid, 1991; 1992a quoted in Jenkin, 1996; 1995)

A range of general criticisms, as well as support, appeared in the pages of Curriculum Perspectives from 1992 to 1998 (ASCA website). CURASS chair Boston (1992, 1993, 1994) explained his view on the development. Collins (1994a, 1994b) and Reid (1992b, 1995) among others, debated the merits of the national curriculum and the profiles.

South Australian based critics (Garrett & Plitz, 1999; Reid, 1991, 1992a, 1992b, 1999; Thomson, 1999; Williams, Johnson, Peters, & Cormack, 1999) were concerned about the constraints put on the curriculum by standardised approaches to design and assessment. These constraints were deemed to apply in approaches that required regular pencil and paper testing as well as in the looser and more flexible, standards approach embedded in the profiles and levels. Reid had concerns that government control, particularly national government control, would have deleterious effects, claiming "educators and school communities are shut out of decision making about the big questions" (Reid, 1999, p. 13). He believed that curriculum was an industrial issue for teachers and the then corporate pressures were to increase control, to define the 'good' teacher as a "skilled technician who can most effectively deliver the expectations set by the curriculum" (Reid, 1999, p. 192) with little professional autonomy and little school autonomy. This concern seems partly at variance with the strongly declared intention of the profile concept to empower teachers, rather than tests, as the adjudicators of student progress (Boomer, cited in Jenkin, 1996).

While other commentators (Lokan, 1997; McGaw, 1994) saw more utility to the design, at least as an assessment support, it is not surprising that there was uncertainty about the classroom use of levels in the minds of teachers.

Critics of 'outcomes based' curriculum approaches added to the complex challenges and possible confusion facing teachers. Donnelly (2007) insists that the Outcome Based Education (OBE) approach followed in the development of the SPFAS was strongly influenced by Spady (1993). Spady was a US advocate of a range of outcomes approaches (Donnelly, 2007, p. 2). Spady, however, is not referenced directly in background documentation as far as the author can ascertain, although the Australia Curriculum Studies Association sponsored a tour by Spady in late 1992 and published his material (Spady, 1993). Certainly the SPFAS was dependent upon clear descriptions of outcomes. Spady's considerations of outcomes may have influenced some teachers and some planners but they

were not the main influences upon the design. As indicated earlier, locally developed concepts (Griffin, 1990; Masters, 1990; Sadler, 1987) were much more powerful influences.

While the profiles were evolving and replacing the attainment levels, South Australia, under the direction of the Brown Liberal government elected in 1993, was also implementing a statewide testing program.

A Parallel Universe - the Testing Approach

A separate initiative was developed in South Australia in the period from 1994 to 1996 and has been repeated annually since then. From 2008 it became part of the *National Assessment Program-Literacy and Numeracy* (*NAPLAN*). This was the introduction of a testing program known as the Basic Skills Testing Program (BSTP) in 1995, after a trial in 41 schools in 1994.

Over a decade earlier Keeves (1982), chair of the Committee of Enquiry into Education in South Australia, had recommended the introduction of revised approaches to student assessment. A concern about student assessment raised by Keeves was the "apparent time consuming nature of such activities in the classroom" (Keeves, 1982, p. 184). The committee argued the benefits of observation schedules among other possibilities to increase the range of skills being assessed. In addition the committee recommended that schools be encouraged "to conduct each year a testing program in the areas of essential skills, numeracy, oracy, reference, problem solving and investigations at the Year 5 and Year 9 levels" (Keeves, 1982, p. 187) but with the decision to implement the assessment to be a local decision for each school. No action was taken to implement the recommendation.

Thus at the point where South Australia was implementing the SPFAS, a parallel development to test students at primary level was introduced, adding to the mix of industrial tension. To introduce a statewide test, South Australia contracted the NSW Department of Education to develop and mark the test in the initial years. The test was conducted at Year 3 and Year 5, and extended to Year 7 in 2001. The declared major purpose of the BSTP was to identify students having difficulties in areas of numeracy and literacy. Each participant was given an individual report indicating items correct and incorrect, graphed in difficulty order, as part of an individual diagnostic analysis. Based on test performance the student was allocated to one of 6 band levels (Hungi, 2003). The initial data collections from 1994 through to 2000 have been extensively analysed by Hungi (2003). The 1997, 1998, 2001 and 2002 waves of the data are part of the analysis reported in this thesis.

The introduction of the test program was controversial. Major concerns were expressed by teachers, while many parents and politicians supported the test program. Hungi (2003) summarises the main arguments. Critics saw the program as unnecessary, not superior to

teacher assessments, and likely to cause teachers to alter their classroom instruction to match the tests and neglect other parts of the curriculum. Supporters saw the program as providing useful feedback, particularly in identifying and assisting weaker students. The program would also assure parents of the quality of the programs in public schools. In 1998 over 95% of the target populations participated in the test program, comparable to other testing years, and confirming for Hungi that parents in the main supported the program (Hungi, 2003, p. 3).

Hungi researched the BSTP data to test the fit of the data to the Rasch model, to explore the item difficulties in Years 3 and 5 and to adjudge whether a common scale could be used for items in both literacy and numeracy for the period 1995 to 2000. He was also interested in the changes in the performance of cohorts of students over time and whether he could quantify growth from Years 3 to Year 5 (Hungi, 2003, p. 6). Hungi's analysis is very detailed and comprehensive. He reports that

> overwhelmingly, the items had adequate fit to the Rasch model and the item means between Grades 3 and 5 compare well year after year. Clearly, the test developers did excellent work in the development of the items and in the allocation of the items to either the Grade 3 or Grade 5 tests. (Hungi, 2003, p. 107)

He also established that the growth in achievement between Years 3 and Year 5 for both numeracy and literacy was consistently about 0.50 logits per year for each of the 6 years in the analysis. He remarks (p 107) that this growth has consistently increased slightly each year, "especially for numeracy". Hungi's detailed analysis of the Basic Skills Test data, re-used in this thesis, provides evidence that the tests are of high quality, fit the Rasch model well, and provide quality reference measures for comparison with teacher's judgements of students involved in both assessment processes. The test data are summarised in Chapter 6.

Summary

The chapter has outlined the general history of the development of the profiles approach adopted in South Australia in the mid 1990s. The model for assessment of student progress was taken directly from the SPFAS, although a parallel attainment levels approach had been developed immediately prior to SPFAS. While the model was implemented in the face of considerable opposition from teachers, this opposition was more about perceived workload than fundamental objections to the assessment model. At the same time as profiles were implemented in SA, primary teachers were objecting to the introduction of tests at Years 3 and 5. A brief synopsis of the beginning of statewide testing was also provided.

Initially, a data collection of teacher assessments of students using the profiles was expected for 1995. A final process was not resolved until mid 1997. Little interest was shown in the refinement of assessments to establish progress within a level. The issue of finer resolution

within a level came to a head in the planning for a collection of student data when teachers refused to provide data unless it was at a finer resolution than a level. A collection process was designed and applied without trial that enabled teachers to indicate the progress a student was making towards achieving the next level beyond the one the teacher believed the student had already achieved. Those data provided by teachers in 1997 and 1998 are analysed in this thesis.

Sadler's concept of standards-based assessment using a process of teacher judgement was a key element of the SPFAS. The next chapter reviews the broad issue of teacher judgement and cites cases from the research literature that illustrate how it has been applied and how it compares to independently obtained measures of student learning.

Chapter 4: Teacher judgement assessment– issues, methods, and case studies

> When we also recall that efforts to achieve high reliability of a test are at the expense of validity, then the balance of advantage falls heavily on the side of using teachers' judgements.
>
> Harlen, 2007b, p. 19

> Record now; teach later.
>
> Clay, 1972, p. 104

This chapter considers research on and examples of teacher judgement assessments. A number of studies are reviewed where teacher judgements assessments are compared to other independent assessments of the same students. However the literature is not rich in investigations of teacher judgement assessment, even though this is a large component of the classroom assessment repertoire of teachers.

The scale or format used to report the assessment is one issue in teacher judgement assessment. Where teacher judgment assessments are compared to test assessments the categories used by teachers to articulate the assessment (a grade, a rating category, a level) are usually fewer than those used to articulate the test assessment (a scale score). While the test scale can usually be regarded as continuous, the teacher judgement scale is usually a small number of ordered categories. The impact of fewer response options for teachers, that is lower resolution relative to the test, is considered. Where teacher and test assessments are not made on equivalent scales the options for comparing results are more limited.

The chapter briefly describes some techniques and issues related to the general comparison of alternative methods of measuring. Using scatter-plots, the alignment of individual teacher's assessments with test assessments for a class of students can be appreciated more easily and criteria can be developed to diagnose whether teacher assessments can be improved. To provide a basis for understanding the degree of difference of alternative assessment processes, the ways in which two quantification processes can match (or mismatch) are considered.

Cases studies from the US, England and Australia are discussed. Assessment strategy descriptors developed from British research provide an insight into the typical behaviours and approaches of teachers as they address how to record data and make assessments in a levels structure. The role of intuition in on-balance teacher judgement assessments is considered.

A synopsis of the comprehensive international reviews by Harlen on the use of teacher judgement, and covering some of the studies in the chapter, provides a consolidation of the value of and potential for teacher judgement assessments.

Issues in comparing teacher judgement and test assessments

Assessment resolution for tests and teachers

Even where test and teacher judgement assessments are reported on notionally similar scales, the precision possible, based on the distance between scale units or the width of ordered categories, can be quite different for each of the two assessment processes. Precision is understood as the combination of the distance between tick marks on a scale, and the relationship of this distance to the smallest increment of learning that a scale (or measurement process) can reliably discriminate.

The implied high precision in test scales can be spurious. Psychometrically developed test assessments are reported on what appear to be continuous scales with scores that represent values on the scales. However, the scale scores are transformations of raw score values using a psychometric model. The original raw scores are a limited set of categories, the number of categories related to the number of items in the test. The highest possible raw score for a given test is a combination of the total number of items and the number of items with part marks awarded. The psychometric model transformation leads to a scale format which then has an apparent greater precision than the raw score increments. The transformed value will most often be either a value to 2 decimal places or its equivalent, through multiplication by 100. Each transformed value is estimated with error, further reducing the effective precision of the estimate. To an uninformed observer fine resolution can be incorrectly inferred for test scales.

On the other hand teacher judgement assessments are routinely made at low resolution, the precision constrained by the structures provided to teachers to articulate their judgement. In the levels structures described in Chapter 3, usually teachers are required to discriminate between increments of the order of 6 to 8 months of learning[14]. In learning terms this is quite low resolution. The levels designers have underestimated the discrimination skills of many teachers.

---

[14] This estimate is based on levels being approximately 2 years of development apart in the current designs. The Australian scheme with the greatest resolution, Victoria (VELS, four categories per curriculum level) therefore has a resolution of 24 months/4 categories = 6 months.

Other category options for teachers range from dichotomous yes/no observations related to specific objectives to various sorts of ordered categories such as A, B, C, ... or 1, 2, 3… , not necessarily related to any underlying developmental dimensions. There are refinements to the categorisation options beyond the broad descriptions above. Marzano (2007) argues for a 4 point rubric system, where an ordered set of outcome possibilities for a topic within any particular grade/Year level is described in increments of 0.5 of a point (Marzano, 2007, p.17-22), providing resolution into nine possible score categories.

Thus a first source of complexity in comparing test and teacher judgement assessments is the resolution of the scales used in the assessments.

Variability in assessment skill and calibration within and between teachers

It should be expected that teacher-test matching is likely to be lower than test-retest matching. Observations of teacher classroom assessment (Dunn, Morgan, O'Reilly & Parry, 2004; Green & Mantz, 2002; Stiggins & Conklin, 1992) indicate there are wide differences in the student behaviours to which teachers attend when assessing students using conventional grading systems. Teacher judgement assessments, particularly when obtained over multiple classrooms and sites as in this current study, are likely to vary between teachers and apriori would be expected to have lower correlations with appropriate tests than test-test correlations designed for a common domain.

The within teacher-test match of assessments for multiple students by an individual teacher is rarely considered in the accessed research studies and statistical reports. An understanding of the ways in which individual teachers match or systematically mismatch test assessments requires multiple assessment cases for each teacher, that is full representation of students from one or more classes. The data explored later in this thesis, and in most of the research in the literature cited, do not include large numbers of replicates of student cases for individual teachers.

Comparisons of teacher judgement assessments with test measures for the same students on the same developmental construct, assume that it is possible for tests and teachers to be quantifying the same underlying construct in approximately the same way. There is evidence (Pedulla, Airasian & Madaus, 1980) that this is likely but that teachers also consider other related variables in their assessments.

Teacher judgement reliability

An issue in teachers' judgement assessments is the possibility of misjudgement at any time, even where the teachers are well calibrated to test scales or any other learning dimensions. Given the higher frequency with which teachers can apply and re-apply regular, simple

assessment processes in the classroom, any error in judging the current learning status of a student will be subject to regular correction, as noted by Shepard (2000): "Classroom assessments do not have to meet the same standard of reliability as external, accountability assessments primarily because no one assessment has as much importance as a one-time accountability test" (Shepard, 2000, p. 67). This thesis accepts the logic of Shepard's observation. Regretfully the repeatability of individual teacher's judgement assessments for specific students, and thus the variability in these judgements, cannot be established from the data analysed in Chapter 7. Few research examples identified in the literature address specifically the variability of assessment-reassessment for individual students by teacher judgement; Rowe and Hill (1996) described later being one exception.

Teacher preparation for assessment

A further difficulty in the design of research on judgement assessment by teachers is the impact of training or preparation in the scale to be used to describe or locate/articulate the judgement. There are two major categories of case studies that involve classroom teachers in making judgements. One set of cases requires an external reference frame, which may be unfamiliar to the teacher. In these cases a lack of experience with the framework or response format might influence the accuracy of judgements.

In the second category are cases where teachers have used a specified framework for an extended period. In these cases teachers are using a framework with which they are familiar, to varying degrees. The tests and the teacher assessments, as a result, might already be in a common framework and use common scales. This is the situation in England's Key Stage assessments and in the Victorian VELS assessments. In principle adoption of these arrangements eliminates some of the potential sources of inaccuracy that unfamiliar response frameworks bring to the assessments.

Lack of research

Given that teacher judgement assessments, explicitly or implicitly, make up a large component of classroom teacher behaviour, it could be assumed that this aspect of teacher behaviour should have led to many studies. It is curious therefore that the veracity of teacher judgments in general, does not appear to be as comprehensively researched as might be expected, even though these judgements contribute importantly to classroom processes. Teachers who do not have good judgement skills would, it is assumed, have great difficulty in targeting support for individual students since they would not understand what was required for each student.

Expectancy research (Hinnant, O'Brien & Ghazarian, 2009; Merton, 1948; Jussim & Eccles, 1995; Rosenthal & Jacobson, 1968; Rosenthal & Rubin, 1978) is one research direction. Here

the hypothesis is that teachers' expectations, exemplified by their judgement of a student's current status and potential, have a strong impact on student success, or lack of success. From this author's perspective the theme of this approach tends to be missing the point. If the inaccuracy of a teacher's judgement contributes to inappropriate outcomes, research on how to improve the accuracy of teachers' assessments is required.

The literature does however have some examples that confirm a fair degree of match of teacher judgements to other independent methods of assessment, particularly to pencil and paper tests in the same general domain. These examples and a small number of statistical summaries from schools systems where teacher judgement assessments are recorded, provide some understanding of the link of teacher assessments to test assessments. These are detailed later.

A further issue considered is the general problem of how two or more methods of assessment are compared. As each teacher is a unique method in the sense of method comparisons potentially independent of any other teacher, the lowest level of the problem is how to compare any specific teacher's judgement assessments and test assessments for that teacher's students.

Methods comparison

Barnhart, Haber and Lin (2007) provide a general overview of approaches to assessing agreement between methods, drawing on the broad range of applications in social, behavioural, physical, biological and medical sciences. The general issue is that of comparing two measures of the same phenomenon, when both are measured with error. Where measurement error is anticipated on both the X and the Y axes, the use of the Ordinary Least Squares (OLS) regression is not appropriate as the result will depend on which of the scales is regressed on the other.

In psychometrics comparing two assessment methods is usually concerned with reliability and validity. Reliability is understood as estimating the degree to which a process (teacher judgement or test) measures the same way each time it is used under the same conditions with the same subjects. Validity is understood as the extent to which the process measures the intended construct.

The comparison of teacher judgement assessments with test assessments is a check of reliability as well as a check of construct validity, particularly if the test is considered as the standard. The reliability and validity of the assessments are confounded. In most research designs of teacher judgement compared with tests (detailed later) there are limited assessment replicates (few students per teacher for alternative 'forms' reliability) and almost no repeats of

the same student for a given teacher (judgement re-judgement reliability), making it impossible to estimate judgement re-judgement reliability for individual teachers. The lack of independence of the assessments of the same student on two or more occasions by a teacher adds a further logical difficulty in teacher judgement re-judgement reliability. Most teacher judgement assessment comparisons in the literature are of aggregates of many teachers' assessments of small numbers of students per teacher, compared with a single test for the same students.

It is assumed the teacher and the test are assessing the same construct. If their assessments match well, the assumption that they are assessing the same construct is supported. However the variation in match of the assessments between teachers and tests may be due to the possibility of many teacher constructs compared with the single test construct. Some teachers may match the test construct and be reliable, some may match the construct and be displaced on the test scale and yet be reliable, confirmed by high correlation coefficients. Some teachers may assess the same construct but be unreliable and finally some teachers may be assessing quite different constructs reliably or unreliably.

The arguments of the thesis support Messick's admonition that performance assessments should be 'construct-driven rather than a task-driven…because the meaning of the construct guides the selection or construction of relevant tasks' (Messick, 1994, p. 22). Messick (1993) argued that the validity of any test depends on whether test results lead to useful, meaningful and fair decisions, thereby making validity a consequence of testing and assessment, introducing the notion of consequential validity. This is consistent with the Fredrickson and Collins (1989) view that subjectivity of scoring, in and of itself, may contribute to the so-called systemic validity of the test. That is, if clear performance standards applied in scoring are also applied by teachers and students in instruction and learning, then subjectively scored tests may "directly reflect and support the development of the aptitudes and traits they are supposed to measure" (p. 28). This systemic validity is seen where program activities enhance test performance and as well the performance of the construct.

In the literature reviewed and in the analyses applied, the methods comparison processes, as a cross-check on validity, can be categorised as belonging to four types: Percentage agreement, Cohen's Kappa (Cohen, 1960), Correlation and 45-degree or identity line comparisons.

The percentage agreement comparison compares categories and reports the agreement values for the same categories in the two assessment methods. Cohen's Kappa extends the comparison. Table 4.1 is based on Altman (1991) indicates one set of descriptors for the categories of agreement. Negative values are possible where the two processes disagree more often than chance.

Table 4.1 Table of Kappa values

Value of *K*	Altman (1991) agreement descriptors
< 0.20	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
0.81 - 1.00	Very good

As the categories on the assessment scale become finer, the data under both methods being compared tend towards continuous scales. In these cases method comparisons based on continuous equal interval scales might be applied. The correlation between cases on two scales indicates the degree to which the two methods order the cases consistently. The Pearson correlation, based on the assumption that the two scales being compared are continuous with equal interval units, provides a first line indication that the scales might be related but offers little more in understanding the relationship (Dunn, 2007). Medium to high correlation, while confirming the scales behave in a related way does not offer a technique to establish the detail of the link between the scales.

The '45 degree line' or identity line comparison is regularly used in method comparisons, including psychometrics. The scatter plot of the results from two methods, assumed to be on a common scale, are plotted, one method on each axis. The points are shown in relation to the line of identity, the line of gradient 1 through the origin.

A close relationship to the identity line indicates a strong link of the two scales. A useful improvement on the process of visual comparison around the identity line is the use of 95% control lines based on the joint measurement error of the two assessment processes being compared as exemplified in Bond and Fox (2007, p. 87). The cases can be converted back to a form of percentage match, using the number of cases within the control lines as a percentage of the total number of cases. Such comparisons require that the error of measurement be established for each assessment result.

One process for this in test assessment is the Rasch model, where the error of measurement of each case is estimated. Where one scale is systematically displaced relative to the other, the scores for one of the measures can be adjusted using the 'average of the differences' method to relate the data points to the identity line. Systematic differences between the two scales can be identified and one of the data sets rescaled so that the scales have common origins. Confidence intervals (95% control lines) as applied in Bond and Fox (2007) are then estimated on the basis of Wright and Stone (1999, p. 65-75), where control lines are set at perpendicular distances from the identity line based on standard error estimates on each scale.

Another statistical process for comparing two methods is orthogonal regression (also described as total least squares (TLS) and error-in-variables modelling). Measurement error on both axes is assumed so that it does not matter which variable is regressed on which (unlike the OLS regression). The TLS regression estimates the line of best fit from orthogonal projections rather than vertical or horizontal projections. The Deming regression, a generalised form of the orthogonal regression, allows the ratio of the variance of the two methods to determine the line of best fit (Dunn, 2007). The Deming regression is used later as a process to establish the systematic scale and gradient differences between some smaller subsets of teachers and the tests.

Approximate model error can be estimated for teacher judgements when teacher judgements on x strands within a learning area can be seen as x independent items assessing the overall learning status in for the student. An estimate of model measurement error is made using the Rasch model. The small number of items, due to the small number of strands, limits the process but it is within the capacity of Winsteps (Linacre, 2006) to fit the strand data to the Rasch model. Teacher judgment assessments can then be brought to approximately the same logit scale as the test for the same learning area. A 45-degree identity line comparison with control lines can then be constructed with the Rasch model estimate of standard error as an adequate estimate of teacher judgement error. This process is applied later in Chapter 8.

Value of the scatter plot techniques

In all the regression/scatter plot techniques an exact match of scales occurs where the slope is 1 (B is 1) and intercept is 0 (A is 0). Any adjustments required to approximate this indicate the extent to which values on the two scales are displaced from each other, i.e. the ways in which the two scales differ are identified. As in the Bond and Fox example above, identifying the systematic displacement (or shift) of one scale relative to the other, provides a potential basis to recalibrate one of the methods relative to the other.

For comparisons of a specific teacher's judgement assessments with test assessments for the same students, a number of ways in which the teachers and test assessments are related can be imagined. A simple exploration of the range of possible relationships of teachers' assessments to test assessments is addressed next, to introduce an understanding of the variety of matches or mismatches that might occur.

Forms of match between independent assessments

Before methods comparisons can be made, data on the scale of one assessment need to be transformed to the scale of the other. Processes to do this are not considered here. Once transformed the degree of match of cases on the two axes can then be established by the relationship of the scatter of the data points to the identity line as described above.

Some general forms of match/mismatch are described below. These relate to the general case for multiple teachers compared to a test as well as the consideration of individual teacher-test relationships, where assessments for all students for individual teachers are considered separately. For the former the described concepts are useful in unfolding the relationship between the scale based on the mean assessments of all teachers and the tests. For the latter, the concepts should influence specifically what might possibly improve matching to the test scale for each teacher.

Based on the scatter plots of the cases assessed by both teacher and test processes, teachers who have any calibration to the same dimension as a test will be identified by their slope and intercept relationship with the test, on the basis of a Deming or Total Least Squares regression. In comparing teacher and test scores through common students three forms of matching, in order of increasing power, are of interest.

Criterion 1- That the data from two sources are in the same order. (High correlation)

Criterion 2- That the data meet criterion 1 and are spaced on the two scales in a similar fashion. (Scatter plot points align with a gradient of 1)

Criterion 3-That the data meet criterion 1 and 2 and that the data points for each student are at exactly the same point on both scales, that is on the identity line subject to joint SEs of the individual estimates. (Scatter plot points align with a gradient (B) of 1 and an intercept (A) of 0)

Meeting criteria 1 and 2 can generate a relatively high correlation coefficient[15] but this might include mismatches on criterion 3. The data can be ordered appropriately, even spaced appropriately but still be displaced from the common scale. In a hypothetical population of teachers a variety of potential mismatches at the individual teacher level can be anticipated. These are listed below. The A and B parameters are assumed to be estimated for a teacher on a Total Least Squares basis.

Mismatch Type 1: *Inability to order students in the same order as the test*; that is the teacher is not able to meet criterion 1. The reason for the mismatch is disagreement on the order. The degree of mismatch may be an indicator of the cause of the disagreement. A few cases out of order would suggest a need for crosschecking the 'not-matching' cases to establish

[15] The issue of scale resolution arises here. In most cases correlation is assumed to be the Pearson product moment correlation even though the teacher assessments scale units (as categories) may be at lower resolution. Depending upon the circumstances, and particularly for level curriculum structures, the teacher scale is assumed to be continuous and equal interval but with readings centred on the midpoint of the level (or sublevel).

whether test measurement inaccuracy, teacher judgement inaccuracy or both might contribute to the mismatch. Mismatch on most or all cases (very low correlation) would suggest the teacher is not calibrated to the test to any useful extent. The teacher is using quite different indicators of where the student is relative to the test. With a low correlation to the test scale the values of A and B will be unhelpful. B will be close to 0.

Mismatch Type 2: *Matching the order but not the spacing*, would indicate an approximate calibration to the scale of the test. Not matching the spacing implies a number of cases will not meet Criterion 3. This would show as a deviation from 1 in B. The regression line for the teacher will probably cross the identity line in the range of interest, indicating that the teacher judgement might be biased above the test for some segments of the test scale and below in others.

Mismatch Type 3: *Matching the order and spacing* (meeting criteria 1 and 2) *but consistently displaced from the test student placement*. This would imply a value of B close to 1, but a value of A quite different from 0. This case would indicate a good general calibration but a consistent displacement of the teacher's perception of where on the scale a student is placed relative to the test.

Mismatch Type 4: *Different resolution detail on one of the scales relative to the other*. A further form of mismatch error can occur where one of the scales for the test or the teacher (the more usual) has fewer categories relative to the other. This circumstance arises where teacher judgements are applied with different unit resolution even though both scales use the same general unit. In a length metaphor this applies where the teacher has a ruler calibrated in metres while the test is calibrated in millimetres. As a result data points on one scale are concentrated at the points representing the degree of resolution for that scale. A stepwise relationship is exhibited where the teacher assessment and the test are well aligned.

Evidence from the literature to be detailed below suggests that many teachers, though clearly not all, can assess students against specified criteria. As a consequence the students are ordered[16] and this order correlates well with other forms of independent assessment.

---

[16] A note on the concept of order. A strict rank-ordering of students for a given learning area, developed normatively on the basis of a teacher's observations, is not a particularly difficult task for teachers. The result is not strongly useful pedagogically as the meaning of the position of each student, in terms of what they know or can do, is not directly revealed. Achieving a similar order on the basis of considering the skills of each student against criteria is a much more useful process as it requires a consideration of the skill profile of each student. However there may be economies for teachers in combining normative ordering with identifying the skill profiles of key students along the rank order,

Teachers ordering students consistent with the order determined by a test is not overly surprising. More significant are the intervals between student placements as discussed above (Criteria 2). If teachers and tests both create approximately similar intervals it can be assumed that they are using similar scales, not just the ability to match orders. In this view the teacher assessor has an understanding of the map and the general length of the journey and the distance travelled so far for each student.

The criteria and the speculation on forms of match/mismatch are developed as part of the consideration of what it means for teacher and test assessments to match. If it can be established that sufficient teachers are generally calibrated to a specific test, the potential exists for improvement in calibration. Even where general calibration can be confirmed, it is assumed a process of moderation will be required to improve the calibration of some teachers and to maintain the calibration of many others. The author conceptualises this problem as keeping A close to 0, and B close to 1.

In further setting the context for the application of teacher assessment it is useful to clarify the processes teachers apply when assessing. Observations of the early stages of the introduction of assessments related to the national curriculum in England in 1991 provide an insight into the transition from a loose assessment process to one related to a standards referenced scheme of the sort advocated by Sadler (1987). Both the transition and the general principles of assessment have parallels with the South Australian situation in the late 1990s. The assessment processes in Victoria which led to data described later in this chapter, are also similar.

Clarifying how teachers make judgement assessments.

Teacher judgement assessment assumes that teachers hold conscious or subconscious hypotheses about each student's learning status. Teachers develop the hypotheses by integrating all their observations for particular students into a judged learning status estimate. A limitation in external observers understanding a teacher's hypothesis is the requirement for the teacher to express the judgement in a form that succinctly describes the status. Some

---

as benchmarks or examples. On this basis, hypotheses about the skill profile of students between benchmark students might help teachers estimate the match to criteria for these students more efficiently.

options for doing this include a scale value from an appropriate test scale or using a scale value related to a levels scale. The latter process is that used in the data section of this thesis. Level values from a level scale are used in the Victorian school system and the Key Stage assessments in England.

When teachers make this assessment what processes do they apply and on what data sources do they draw?

In the early stages of the introduction of teacher judgement assessment (TA) in the England national curriculum Gipps, Brown, McCullum, and McAlister (1995) observed the strategies teachers used in teacher judgement assessment. Gipps et al. developed a descriptive classification that typified the wide range of teacher behaviours they observed and explored in interviews, as teachers made their initial public judgements of student learning status.

Teachers applied one of three major strategies when they were required to provide a summative assessment in a levels framework. These three strategies for teacher assessment of students were categorised as intuitive, evidence gathering, and systematic planning.

Teachers using the intuitive strategy made a "kind of gut reaction" judgement (Gipps et al., 1995, p. 36) based on their memory of what the students could do. As a result it was difficult to observe any ongoing teacher assessment support processes such as record keeping, assessment focused events or conversations.

Teachers using the evidence gathering strategy gained as much evidence as they could and become hoarders who kept everything. Gipps et al. indicate that these teachers preferred not to rely on memory because the number of elements to be assessed was too great. They planned assessment at the same time as they planned their topic work. One motivation for evidence gatherers appears to be self-protection in case they are challenged, indicating perhaps less confidence in their processes. Though not described as such by Gipps et al., this strategy also implies a concern with the detail rather than the bigger picture.

The systematic planning strategists planned assessments on a much more systematic basis that became part of their practice. They usually committed all the detail of the assessment schemes to memory, but also had at hand reference documents. This fits with the assumed strategy expected by Thorndike in the application of his handwriting and prose scales (Chapter 2). The systematic planners believed strongly in ongoing formative assessment, which usually involved note taking about specific students. Apparently they distrusted relying on memory for keeping records about students. Taking advantage of the openness of the levels scale, they were willing to assess children on higher levels without necessarily having taught the content first. Assessment became a learning process for the teachers. They

distilled attainment from all other information and did not confuse it with attitudes, context or biographical data.

Based on this broad analysis, the repertoire of assessment approaches in SA (and mostly everywhere it is assumed) would approximate the range identified by Gipps et al.. In the space of the two years of observations by Gipps et al., the strategies moved in the direction of increasing the proportion of systematic planners as one outcome of the new national reporting requirements. Teachers' assessments at the start were mainly intuitive, and while some teachers still made intuitive judgements (as defined by Gipps et al.) after two years, many more teachers were basing their judgements firmly on documented evidence. This observation begs the question of whether the observed change to systematic planning strategy was any less intuitive. Teachers had developed refined approaches to their observations and recording, and had added and become very familiar with the organised but still ambiguous reference frames. This author suggests that intuition was still part of the judgement process, a little like the internalisation of scales and standards expected by Thorndike.

Gipps et al. advocate the more systematic, evidence based techniques of systematic planners. The position taken in this thesis is that recording is important and might be achieved in a shorthand fashion using the judged scale position. This would be consistent with the planning of the systematic planners, where strategies to integrate all information both recorded and recently remembered are of value. The judgement processes of the expertly prepared systematic planners might also become intuitive, given the efficiency with which the judgment might be made. What distinguishes the intuition of the systematic planner from the initial intuitive assessor is the store of internal reference frames, the evidence considered and the developed skill in articulating the judgement. This position is consistent with those of Klein (1999, 2009) on expert decision makers and Sadler (1987) on connoisseurship. Intuition, according to Klein "depends on the use of experience to recognize key patterns that indicate the dynamics of the situation" (Klein, 1999, p. 31). What typifies intuition is the speed with which a judgement is made (Gladwell, 2005; Klein, 1999).

Intuitive decision makers, under the Klein conception, draw on patterns, anomalies, understandings of how things work, likely preceding and post events, and their ability to discriminate pattern differences that are very small (Klein, 1999, p. 148-149). It is this ability to "see the invisible" (Klein, 1999, p. 147) when experts make judgements that provides support to the main proposition of this thesis. This is the hypothesised skill of expert teachers, using frameworks to efficiently and accurately judge and record the learning status of a student. Intuitive decision makers feel uncomfortable about "trusting a source of power that seems so accidental" (Klein, 1999, p. 31). Intuitive assessors are often unable to describe how they made their judgement and as a result often find justifying their decision difficult.

Reliance on the intuitive expert opinion can be seen as in conflict with evidence-based assessment, particularly where the experts are themselves uncertain about how they came to a decision. The proof or otherwise of expert opinion, however, is in how well it matches the results of other assessment processes.

Having set the scene in terms of the typical range of processes adopted by teachers, what have investigations of comparisons of teacher assessments with independent assessments shown?

Studies/examples of the use of teacher judgement in research and classroom practice

The bulk of the rest of the chapter considers the application of teacher judgement in research studies and in the general structure of assessment processes in three assessment cultures. These are in the US, England and Australia. The examples illustrate the variety of ways teacher judgement has been researched and cases where teacher judgement has been formally applied in school systems. The essence of this review is to establish a view of the quality of teacher judgement. Other countries, Canada as an example, are not included although teacher judgement is part of the assessment process in some provinces (Ministry of Education, Québec, 2002). Scotland and New Zealand are also not included due to space limitations, although they also provide examples of how teacher judgement assessment has been applied.

The US experience is described, in the main, from a synthesis of the US research and the main findings from that, rather than from all individual cases. A small number of case studies illustrate the methodologies applied. The England experience is based on the implementation of a teacher judgement component for assessment in the national curriculum and the trends in this component over a series of years compared to the tests for the same learning areas. The England school system had, for over a decade, ongoing parallel assessments by teachers and tests at all Key Stages. These parallel assessment processes are under review. Testing (SATs) at Key Stage 1 was abolished in 2004. Teacher assessments only, to a strict protocol, have applied at Key Stage 1 since 2005, meaning the potential to compare assessments has disappeared.

The Australian examples illustrate some cases where teacher judgement has been applied. Victoria is the state where teacher judgement has been most used in classroom assessment and in school system reporting. At Years 3, 5, 7 and 9 it is possible to compare teacher judgement assessment with a test assessment but little documentation of the comparisons is available. Teacher judgement is the major part of upper secondary assessment in Queensland and the Australian Capital Territory. While both systems are excellent examples of systems confidently relying on teacher judgement assessments, they are not treated in detail as neither offers test data as a cross-check for validity. Western Australia and New South Wales are not treated even though comprehensive testing arrangements apply for the opposite reason: no

teacher judgement assessment data are collected. One South Australian case study is considered in this chapter, followed by more analyses of South Australian teacher judgement assessments in subsequent chapters.

US research

Two different interests affect the focus of US research.

One set of interests is concerned with the expectancy effect of teachers (Rosenthal & Jacobson, 1968), i.e., if teachers expect pupils to do well, then they are more likely to do so. This has a parallel with the self-fulfilling prophecy, of Merton (1948), where the beliefs teachers hold about students lead to their fulfilment. Any potential bias of the teacher in their judgement of a student, it is argued in this view, will influence the self-image and longer-term development of the student. The expectancy effect, when an inaccurate judgment occurs, can have a positive effect (if overestimated) or negative effect (if underestimated). There is dispute about the size of the effect (Jussim & Eccles, 1995; Rosenthal & Rubin, 1978). The likelihood of teacher judgement inaccuracy is not disputed here nor is the notion that some subsets of students are affected by estimation errors. Hinnant, O'Brien and Ghazarian (2009, p. 69) establish that the reading of minority boys had "the lowest performance when their abilities were underestimated and the greatest gains when their abilities were overestimated". This thesis argues that if teacher judgements do have implications, including the expectancy effect, understanding how accurate these judgements are and how susceptible they are to improvement in accuracy, is important.

The second major US research interest is that of the current accuracy of teacher judgments assessments. Research on the accuracy of teacher judgements is reviewed by Hoge and Coladarci (1989), and Perry and Meisels (1996). The latter review was initiated as a basis for considering the options for data collection for the Early Childhood Longitudinal Study of the National Center for Educational Statistics.

Hoge and Coladarci (1989) reviewed a number of correlation studies and identified two major subcategories for the studies; 'direct' and 'indirect' assessments. Direct teacher judgements require an explicit link between criterion and judgement. In the indirect approach the teacher is given little guidance as to the nature of the construct. Unsurprisingly, the median correlation in 16 studies reviewed was higher for direct assessments than for indirect (0.69 versus 0.62), indicating, from the authors' perspectives, the value of making the construct explicit to improve the quality of teacher judgements.

Perry and Meisels (1996) provide a wide-ranging review of what the research indicates about the accuracy of teacher judgements of students' academic performance. In addition to the direct/indirect dichotomy of Hoge and Coladarci, Perry and Meisels identify specificity, norm

and criterion referencing as issues. They conclude that the more direct and the more specific the judgement the greater the accuracy and the consistency of the judgements made. They also find through comparison of criterion-referenced measures with specific standards, that criterion-referenced measures provide greater consistency than norm-referenced measures. The accuracy of norm-referenced judgments is dependent upon the teacher's familiarity with the reference group, which for reference groups beyond their own class proves to be more difficult (p. 11).

They also establish that accuracy is dependent upon the domain in which the judgement is to be made. Citing Coladarci (1986) they find that assessments of reading and mathematics are more accurate than in science or social studies, partly they speculate, due to the degree of observability of the learning. Activities that are concrete (reading aloud, worked mathematics examples) allow teachers to collect more evidence for their judgements (p. 12-13).

The accuracy of teacher judgement is influenced, according to Wasik and Loven (1980 cited by Perry & Meisels, 1996), by the number of categories teachers are required to discriminate and the phenomenon of observer drift. Observer drift occurs when categories are interpreted differently or are not seen as clear and distinct. Perry and Meisels find evidence for individual teachers' judgements being consistent over time. Variability of judgements across teachers is also observed (Perry & Meisels, 1996, p. 17). They also provide evidence for improvement through training. This evidence is based on Meisels, Liaw, Dorfman, and Nelson (1995) where trained raters showed high inter-rater reliabilities, while raters compared to untrained teachers showed a lower reliability (0.88 versus 0.68). Although the conclusion is not drawn directly, the likelihood of teacher judgement accuracy improving with training and feedback is high (Meisels, Bickel, Nicholson, Xue & Atkins-Burnett, 2001).

Perry and Meisels acknowledge the issue of bias, the concern of the expectancy effect researchers, as occurring but to a lesser extent and usually in understandable circumstances. The major bias reported by Perry and Meisels is for teachers to be less able to estimate the skills of less successful students with a bias towards better accuracy with successful students (p. 20).

Coladarci (1986) found that aggregate scores of teachers' judgments of their students' responses on achievement tests correlated positively and substantially with aggregate scores of students' actual responses. Teachers accurately judged their students' responses to individual items for approximately three quarters of the total number of test items; but the accuracy of teachers' judgments varied significantly by subtest. The test-teacher correlation over all students, by subtest, ranged from 0.67 to 0.85. This study is one of the few that has adequate replicates of judgement per teacher to consider individual variation in teachers'

judgement accuracy. The study confirms some degree of individual differences among teachers in the accuracy of their judgments. In some cases teachers were able to predict up to 95 to 100 percent of the student responses, helped by the fact that high performing students were easier to estimate. However, based on a one-way analysis of variance of the success rates, there were only small differences in teacher ability to judge students' scores.

Teachers were least accurate in judging low-performing students and most accurate in judging high-performing students (Coladarci, 1986). In making judgments for a moderate or low-achieving student, there were many items that the student could not answer correctly. These results point tentatively to the implication that students who perhaps are in the greatest need of accurate appraisals made by personalised judgement of the teacher, are precisely those students whose current learning position has a greater chance of being misjudged. The study confirms however that teachers were competent estimators of student's scores.

There are also concerns about gender bias and negative assessments, particularly of low SES boys. This concern interacts with the issue of good classroom behaviour versus poor behaviour. Perry and Meisels find that "while some teachers' judgements may reflect bias of one sort or another, teachers as a group base their judgements of students' academic performance on their knowledge of students' academic skills" (Perry & Meisels, 1996, p. 24).

Wright and Wiese (1988) establish that teacher judgements correlate well with test results. They show that teachers' ratings of student achievement, when deliberately isolated from effort which teachers often compound into grades, correlate more highly with SRA test scores than with the teachers' original grades. They speculate that test scores indicate learning and that teacher grades indicate performance.

Demaray and Elliott (1998) examined differences in teacher accuracy as a function of using similar versus dissimilar judgment indicators. Item predictions on standardised achievement tests produced higher correlations than those found by rating scales, adding support to the Coladarci (1986) finding that the use of similar (direct) over dissimilar (indirect) indicators led to higher correlations of teacher judgements with student performance.

Fuller (2000) considered the ability of teachers to predict the likelihood of students passing the Ohio Fourth or Sixth Grade proficiency tests. A very limited category scale was used ('likely to pass', 'uncertain to pass' or 'unlikely to pass') to predict three months in advance of the tests, teachers' view on the likely category the student would be in. Ninety teachers were involved over 23 schools. The median efficiency was 67% correctly assigned for passing in science and 81% in mathematics. Predicting those who were unlikely to pass was 39% correctly allocated for mathematics and 54% for science. The design was restricted in its potential to establish how refined teachers' predictions could be through the use of the

pass/fail/uncertain categories only. Methodologically the benefit of estimating a scale position for each student is highlighted by default. In the absence of a common scale across teachers the ability of teachers to express how they see the progress of each student is severely limited.

Using teacher judgement estimates where direct quantification is feasible highlights another complication in researchers appreciating where teacher judgement might be most appropriate. Feinberg and Shapiro (2003) required teachers to make estimates of readily quantifiable skills, words read per minute, rather than have them to count them directly. It is consistent with the line of this thesis that effective teachers should be able to make reasonable estimates but the need for professional inference ability is less when the behaviour or skill is readily observed directly. Estimating a value that can be obtained directly and accurately in a minute or so is introducing professional judgement unnecessarily. Estimating test scores on the other hand, or more usefully scale values (as distinct from raw scores), for students has utility if the data are generated in a few seconds as against many weeks for off-site scored tests. The requirement to estimate is even less useful where the teacher is not familiar with the data attribute and has no experience of using it in the classroom, as applied in this study.

US research conclusions

In summary the US cases confirm moderate correlation between teacher estimates of students' scores or rankings, though most studies do not have a design where the teacher scale and the test scale are in, or converted to, the same scale units. The correspondence is greatest when assessments are direct and use approximately similar units or where teachers estimate which test items are likely to be achieved by individual students. Generally, the opportunity for US teachers to be shown to be effective on-balance assessors of students learning status is inhibited by the assessment conventions that routinely apply. Although descriptions of standards to be met at particular grades are now commonplace in US school districts, the research literature is light on independent teacher judgement assessment estimations of students compared to the tests now generally required from Grades 3 to 8.

There are a number of inadequacies of the US research. Most studies are one off, without addressing the improvement in judgment accuracy that might come with multiple repeats using the same teachers over two or three years. Very few studies address the variability of judgement accuracy across teachers, taking instead very small samples of students per teacher. Teacher judgement is treated in aggregation rather than as a skill that might vary significantly across individual teachers.

Perry and Meisels query whether enough care is taken in the choice of the criteria against which teacher judgements are evaluated (Perry & Meisels, 1996, p. 27) and the degree of

understanding teachers have of the assessment they are asked to make. Much of the US research is made complex by the lack of common teacher and test scales.

Perry and Meisels criticise the lack of acknowledgement that teachers "because they observe and interact with their students on a daily basis may be in the best position to make judgments about them" (1996, p. 27). Meisels, Dorfman and Steele (1994) clarify the meaning of standardised. They point out that standardisation is not limited to standard scores or norms. They see standardisation as "formal rules of operation and explicit principles of interpretation [that have] been studied sufficiently to understand how different groups of children, in different situations, will react to a particular assessment" (Meisels et al., 1994, p. 204). Under this definition a wide range of assessment activities is possible, including reference to empirically developed progress maps as standardised approaches.

US research is limited in clarifying the accuracy of teacher judgement assessments. On balance the general impression is that teachers are adequate, if mixed, in the accuracy of their judgements. Most investigations are one-off with limited consideration of techniques to improve the quality of judgements.

Teacher Assessment in England

In the early 1990s teacher judgement assessments in England (teacher assessments -TA) were used in ways similar to that advocated in the Statements and Profiles for Australian Schools (SPFAS). TA was used as part of the summative assessment process of the Key Stages (KSs) of the national curriculum. There are stronger similarities in the Australian approach to assessment by teachers with the England approach than with the US cases above. Many of the same issues that applied in Australia arose, including the issue of lack of subdivision within a level (National Curriculum Council, 1991; Daugherty, 1997).

While teachers report teacher judgment assessments at Key Stages 1, 2 and 3, direct comparisons with the tests at the same stages are rare. This is particularly true for matched individual student and individual teacher comparisons, where only limited comparisons are reported. This lack of direct comparison of teacher assessments and test assessments appears to be a missing feature of the England research into teacher assessment.

Part of the reason for not comparing the actual teacher assessments and test data for individual students may be a lack of confidence in the quality of the test assessments (Stobard, 2001; Tymms, 2004). These concerns are summarised briefly in Appendix 3. The Appendix indicates that using the test data, as the assumed best possible independent estimate of a student's developmental position on the levels scale is problematic. Mismatch of teacher and test data would not necessarily indicate inaccuracy in the teacher assessment but may reflect inadequacy in the test data analysis, even given the broadness of the level scale.

However being aware of the patterns of the two assessment processes over time and the persistence of these patterns by subject provides some hints as to their relationship.

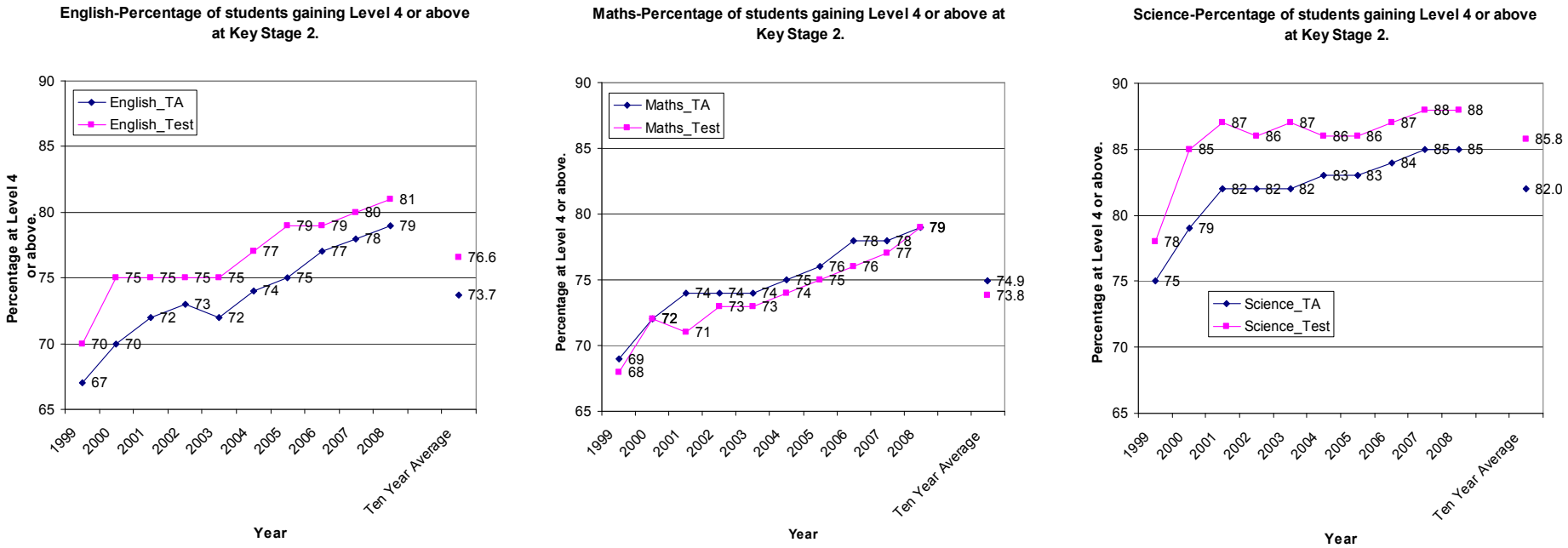
General trends in student assessments compared with test assessments: England 1999 to 2008

Figure 4.1 shows the trends in teacher and test data in the England national assessment at KS2 from 1999 to 2008. Trends are reported as the percentage of assessments at level 4 or higher, that is, students assessed as being below level 4 are not included. The comparisons of test and teacher assessments are for England in aggregate. The assessments are not matched at the individual student level. However the same student population is teacher assessed and test assessed.

For English language, the percentage of students identified by the tests as being at level 4 or above has consistently been greater than the percentage identified by teacher assessment. The average difference for the ten-year period has been approximately three percentage points, but the two data sources have tracked each other consistently over the full period. The differences between the lines are approaching the possible error of measurement related to a one-mark difference in the placement of the level boundaries (Tymms, 2004; c.f. Appendix 3).

Were the two English trajectories essentially identical they would be expected to crisscross with error accounting for the differences. That the percentages of students identified by teachers as being above a particular level are consistently lower than those identified by the tests suggests that teachers' estimates of the position of level boundaries, on average, are higher on the level scale than the test derived cut points. Based on the general matching concepts identified earlier in the chapter, English teachers at KS2 could be assumed to be well calibrated, on average, to the test scale but consistently displaced, applying slightly more severe criteria. This displacement hypothesis assumes a pattern of relationship at an individual teacher level for which there are no data publicly available to enable further exploration. As will be shown later there are broad indicators of the degree to which teacher and test assessments match for individual students but no data to help confirm that individual teachers assess consistently. There are no data to show that individual teacher assessments are consistently above, below or the same, relative to the test criteria for level boundaries. The relationship of teacher assessments to test assessments over all teachers, suggests that a number of teachers must be calibrated to the test scale but displaced up the level scale, for the pattern to persist.

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



Sources: Statistical First Releases, Department for Children, Schools and Families, UK

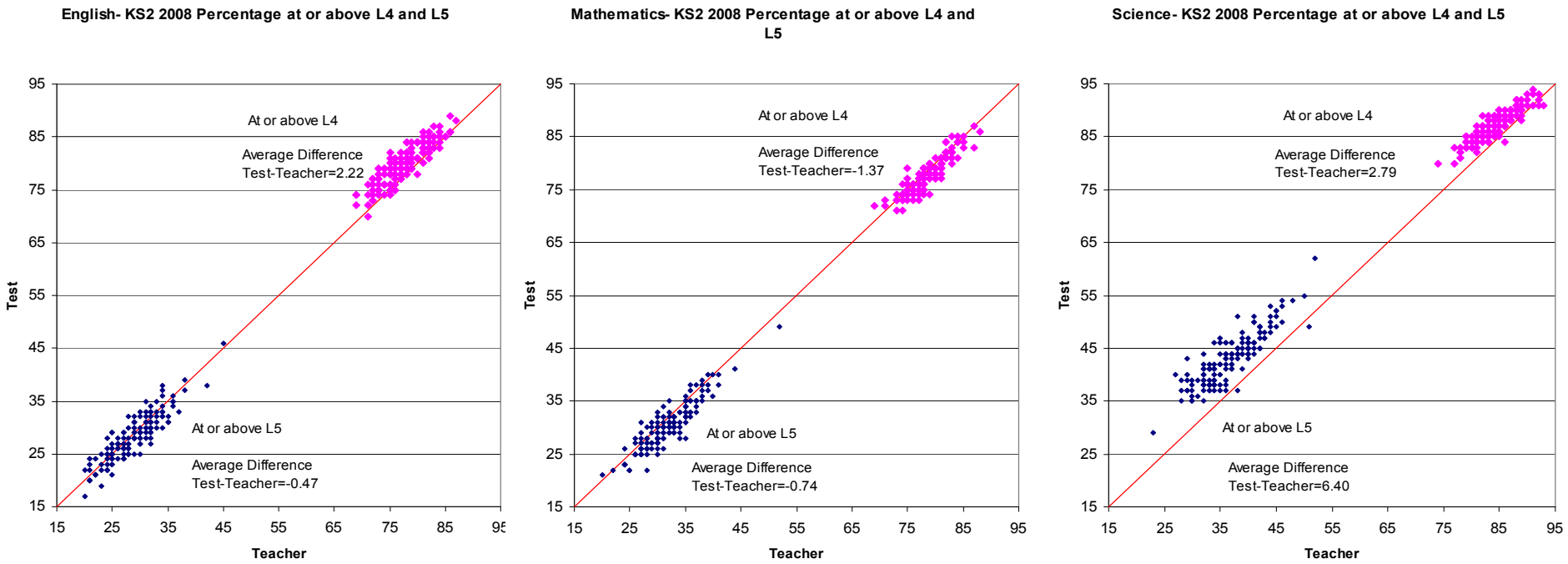
1999-SFR29/1999, 2000-SFR43/2000, 2001-SFR37/2001, 2002-SFR 21/2002, 2003-SFR 20/2003, 2004-SFR 30/2004, 2005-SFR 31/2005, 2006-SFR 31/2006, 2007-SFR 24/2007, 2008-SFR19/2008, SFR 06/2009,

The patterns for mathematics show similar consistency, except that teachers estimate slightly higher percentages of students at level 4 or above relative to the test, the reverse of the English language situation. Again teachers and tests vary by about one percentage point on average over an extended period. In science the patterns of tracking of teacher and test assessments appear parallel as for the other subjects, although the gap narrowed in the period from 2000 to 2006. As for the English language results, the teachers' scale is displaced implying that the average perception of teachers for level boundaries places them slightly higher relative to the test cut points for levels.

The percentage of students at or above a particular level is however a very general criterion for comparing the relative effects of test and teachers assessments. Given the uncertainties at the boundaries for both the test and teacher allocations to a given level, the maintenance of consistent patterns over an extended period, including the close shadowing of the general improvement trends over time (even though displaced), suggest a strong link between the teacher judgement and the test assessment.

A second view of the relationship can be observed through the Local Authority (LA) tables of the annual Key Stage reports (Figure 4.2). Here the data are matched at local authority level, but teacher and tests assessment are still not compared for individual students. The advantage of these plots is that some of the variability in the assessments by geographical location (and thus socio-economic status) is highlighted. The within-LA, school and teacher variability remain masked. Figure 4.2 plots the test and teacher summaries from Tables 6 and 7 of the *National Curriculum Assessments at Key Stage 2 in England, DCSF (2008) report (SFR 20/2008*). From these tables the average percentage per LA for teacher and test assessment is plotted independently for the students at or above level 4, and those at or above level 5. For English language the L4 and above plots sit mainly above the identity line (average difference +2.2 percentage points). The L5 and above group are more evenly spread around the identity line (average difference –0.47 percentage points). The L4 difference is consistent with that displayed in Figure 4.1 (test above teacher). Implied in the result for L5 is that the teacher and test placements of the L4/L5 threshold are closer than for L3/L4 threshold.

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



Children, Schools and Families, downloaded from http://www.dcsf.gov.uk/rsgateway/DB/SFR/s000836/index.shtml 3 April 2009.

For mathematics and science the patterns for L4 are consistent with those in Figure 4.1 (mathematics-teacher above test, science the reverse). Overall the spread of the points along the identity lines indicate that teachers' judgements within each LA are similar to the test assessments for the same LA student samples.

Figure 4.2 illustrates the wide spread of student achievement across LAs; about 20 percentage points from the lowest to the highest in each subject. This spread in the distribution of performance by LA shows much greater diversity than the difference between teacher and test assessments in any particular LA. The teacher and test assessments are close, on average, for each LA. The patterns of relationship between teacher and test assessments indicate the same general consistency by location as shown for calendar year in Figure 4.1. Specific patterns apply for particular subjects, suggesting that part of the variation between teachers and tests relates to the calibrations of the test and teacher assessment to the levels scale specific to each subject.

There is likely to be a variation in the degree of match of teachers to tests when the assessments of individual students are considered. A hint of the size of this variation in assessments can be obtained from the few cases where assessments for individual students have been compared.

What matched data for individual students from three sources say about teacher and test assessments

England has the richest data for comparing the match of teacher judgements to test assessments. The data are reported, however, at very low resolution. The assessments are generally reported at a KS level, or in 1/3rd of a level in some cases. These data have been available at an individual student level since the early 1990s but do not appear to have been publicly or officially analysed for degree of match very often at either student or teacher level.

Data are reported annually at a national, local authority and school level. At KS2 and KS3, aggregate data from teacher assessments and tests assessments are presented side by side and they are summarised independently without exploring the degree of match at an individual student level (Statistical First Releases, 20/2008 & 06/2009). At KS1, assessments prior to 2005 required general teacher judgement assessments as well as standardised teacher-managed assessments. Since 2005 only the teacher judgements have been reported (Statistical First Release, 21/2008).

That these data have not been analysed by official entities was confirmed by the answer to a parliamentary question in February 2009 from the Minister of State for Schools and Learners (Knight).

> The Department has not made an assessment of the level of agreement between teacher assessment and key stage test results at key stages 1, 2 and 4. This is an area being considered by the Expert Group on assessment. Internal analysis of the level of agreement between the 2007 key stage 3 (KS3) teacher assessments and national curriculum tests has been undertaken … Analysis of these data indicates that there is a reasonable match between test performance and teacher assessment data. Where there is not, the teacher assessments are equally likely to be higher or lower than the performance test level achieved. (Knight, 2009)

Three data analyses, in which the direct matches of teacher and test assessment for individual students are made, are summarised below. The first case is a five-year analysis of data for the Worcestershire Local Education Authority by Durant (2003). The second case was part of an evaluation of Key Stage One changes (Assessment and Evaluation Unit, 2004). The third case is derived from the answer to the parliamentary question above (Knight, 2009). Taken together the three sources provide an indication of the degree of match between teacher assessments and test results at the individual student level.

Source 1-Five years of data- Worcestershire Local Education Authority

Durant (2003) reports the degree of match of teacher and test assessments for 5 successive years, from the 1997/1998 school year to the 2001/2002 school year for KSs 1, 2 and 3. The data were extracted from the administrative records provided to the authority from the then Department for Education and Science (DfES), suggesting that it is likely that other authorities have conducted similar analyses. Durant's analysis seems to be the only one that has been reported publicly.

Tables 4.2 and 4.3 are derived from Durant (2003). Table 4.2 shows a grand average summary of 5 years of data, ranging from 27,000 cases at KS3 to 45,000 cases at KS2. The table reports the percentage of cases where the teacher and test produce a match, ranging from 96% at KS1 to 53% at KS3. In this view the data are reported as level categories. In almost all cases of not-matching the mismatch is by one level only. Given the large size of the KS level as a unit (equivalent to two years of development) matches should be close and mismatches should be confined to adjacent categories. As noted earlier, the high match at KS1 is partly because the assessments are not independent; the teacher applies and marks the tests/tasks for the assessment.

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.

	KS1		KS2			KS3			
	Writing	**Reading**	**Maths**	**English**	**Maths**	**Science**	**English**	**Maths**	**Science**
Test above Teacher	2%	5%	6%	15%	9%	17%	22%	18%	21%
Matched Cases	**96%**	**93%**	**90%**	**76%**	**79%**	**74%**	**53%**	**69%**	**63%**
Teacher above test	3%	2%	3%	8%	11%	9%	24%	13%	16%
No of cases.	32578	32559	32651	45135	45102	44854	27632	27646	27562

Source:
Derived from Durant (2003), Annexes 1-9

Table 4.2 illustrates that matches diminish as the stage of assessment increases. KS1 has the highest percentage of matched cases, possibly due to the lack of independence of the assessment processes. Matches at KS2 are around 75%, with the mismatch being in the direction of the test assigning a higher level than the teacher in English and science. Evidence presented earlier in the chapter indicating systematic and consistent differences in the scales over time by subject, suggests that the teacher and test scales are consistently displaced from each other for these subjects.

KS3 data indicate a wider variation in matched cases, at 53% for English and 69% for mathematics. Mismatches are spread evenly for all subjects, about 20% of test assessments above teachers' and approximately 20% teachers' assessments above test. As discussed elsewhere (Appendix 3), the setting of cut points for the level boundaries for the test and the inherent measurement error for all test assessments influence which individuals sit either side of the boundaries. While this has consequences when the assessment of the individual is considered, the impact of measurement error on who sits either side of the cut point has a negligible effect on the degree of match, assuming the measurement error is random.

Table 4.3 provides a more refined view of the KS1 data. At level 2, where more than 60% of the cases sit, teachers place students into categories equal to one third of a level. From the Table 4.2 view, more than 90% of the cases match but when the data are placed into thirds of a level for level 2 (Table 4.3), the direct match is reduced to between 58% and 49%. However between another 30% and 40% of cases are within 1/3 of a level of a match, with 6% to 14% of cases within 2/3 of a level of matching.

Table 4.3 Matches of Grand Average Teacher and Test Assessments-Worcestershire LEA, 1997 to 2001, with 1/3 level as unit of reporting

	KS1		
	Writing	**Reading**	**Maths**
Test above Teacher (1 level)	2%	5%	6%
Test above Teacher (2/3 level)	4%	10%	2%
Test above Teacher (1/3 level)	14%	20%	18%
Matched Cases	**49%**	**58%**	**50%**
Teacher above test (1/3 level)	25%	10%	16%
Teacher above test (2/3 level)	4%	4%	4%
Teacher above test (1 level)	3%	2%	3%
No of cases.	**32578**	**32559**	**32651**

Source:
Derived from Durant (2003), Annex 1-3

Durant concludes that there is "not much" difference between teacher and test assessed levels. He wonders whether teachers might have been influenced by reviewing test scripts but concludes that if this had been a factor the match would have been greater than that recorded (p. 6).

Source 2- Key Stage One (7 year olds) revision 2004

An evaluation report of a trial of using TA only at KS1 (Assessment and Evaluation Unit, University of Leeds, 2004) considered the degree of match, as did Durant (2003). For a random selection of schools, covering approximately 3000 students, the direct match of teacher and test/task assessments was approximately 90%, comparable to the level match summary (Table 4.2) and a lot higher than the Durant (2003) summary using the division of level 2 into three subdivisions (Table 4.3 above).

Taking Reading as an example (Assessment and Evaluation Unit, University of Leeds, 2004, Table 2.16, p. 39; Table 2.20, p. 41) 89% of assessments were identical in 2003 and 90% in 2004 with about 9% of the remaining cases within 2/3rds of a level for 2004. Cohen's Kappa values (Table 2.24, p. 42) compare 2004 assessment match rates to the 2003 rates across all three subjects assessed. Values ranged from 0.74 to 0.81 in 2003, to 0.89 to 0.91 in 2004. On the basis of the translation of the Kappa values to descriptions, this was an improvement in degree of match from 'good' to 'very good' (c.f. Altman, 1991) or in the terminology used by the Evaluation team, from 'substantial' to 'almost perfect' (c.f. Landis & Koch, 1977). There is an indication in the evaluation of a possible contamination impact of tests/tasks on the teacher judgement assessments in 2004. This was a result of Teacher Assessment being "required to be informed by the task/test result and therefore is not independent" (p. 38). The more independent data for 2003 confirm, however, that both forms of assessment match well.

Source 3- Key Stage Three – Teacher/Test comparison 2007

In response to a parliamentary question (Knight, 2009) the Minister of State for Schools and Learners provided data on the match of teacher and tests assessments for individual assessments and confirmed that this view of the results is not regularly reported. For English at KS 3, a grand average percentage match, weighted in proportion to the number of students in each cell, is 61.5% of cases matching exactly, a higher match than Table 4.2 (Durant, 2003) where the match was 53%.

In mathematics the grand average of the matches is 70.3%, which compares well with the Durant data (Table 4.2) where the matching rate for mathematics at Key Stage 3 is 69%. The mismatches are spread evenly above and below the matching zone. Almost all cases are within one level. For the science data the estimated grand average match of about 65.3% compares favourable with Durant's 63% for science. The mismatches are distributed evenly above and below the zone of exact match.

Summary of Teacher Judgement in England

Assessments of students on the basis of teacher judgement have applied in varying forms since the introduction of the national curriculum in 1990. Up until 2005 the Key Stage 1 assessment required both teacher and standardised assessments for students at the end of Year 2. Since 2005 assessments at Key Stage 1 have been by teacher judgement only but to a specified protocol.

Key Stages 2 and 3 ran, up to 2008, parallel assessment processes. Students have been tested near the end of Year 6 and Year 9. The aggregate percentage figures (students at or above specific levels) by subject for teacher and test assessments are very similar, varying in recent years by between 0 to 3 percentage points. As a result of this proximity, the assessments from both sources have tracked together as the percentages of students meeting or exceeding level thresholds have increased, notwithstanding that Tymms (2004) challenges the comparability of the results over successive years, particularly prior to 2000.

When individual student assessments are compared, the degree of match between teacher and test assessments diminishes as the key stage increases, from about 90% at KS1, to c. 70% at KS2 and around 60% at KS3.

KS 1 teacher and test based assessments matched very closely up to 2004. Since 2004 only teacher judgement data are reported. In the cases of mismatch, the mismatch is almost always by one level only. This is to be expected given the wide range of learning described in any given level.

The assessment skills of individual teachers cannot be established from the data sources reported, limiting any consideration of whether differences between teacher and test assessments apply to all teachers or to a smaller set. The literature does not report the patterns of match for individual teachers at the school level. Based on the mandatory reporting to parents, teachers themselves (and parents of each student) are most likely to be aware of the degree of match of their assessment to test assessments. As far as can be determined patterns of match, reflecting systematically displaced teacher and test scales, as hypothesised earlier in this chapter, have not been explored. The impact of teachers reflecting upon the match of their assessments with tests assessments, that is whether teachers increase their degree of match after feedback, appear similarly not to be widely reported. That the general patterns from both assessment sources are very similar suggests some moderation processes apply. Recent testing difficulties (Sutherland, 2009) and pressure from teachers (Garner, 2009, April 12) suggest that testing is under consideration for removal. If tests were removed, England's schools would move to the same position as those in Wales and Scotland where reporting is solely by teacher judgement. Given the large unit size (i.e. low-resolution scales involved and the summative only nature of the assessments) this might not be problematic, although one source of potential feedback to teachers and general moderation information would be lost.

Approaches to teacher judgement in Australia

In Australia teacher judgement assessment applications include national assessments of language development, a state school system that has required the reporting of teacher judgement assessments for over a decade, research projects that have used teacher judged profiles to monitor student learning development and two state systems that have used teacher judgement as the major assessment process for the end of Year 12 certification. Student assessment based on teacher judgement has thrived in Australia. Some initiatives, the Statements and Profiles for Australian Schools (SPFAS) in particular, have not met the expectations of their developers. Initiatives based on the SPFAS concept of levelled learning descriptions however have been applied in the Victorian state school system for more than a decade. Teachers there have been required to keep records and to report to parents using teacher judged levels. Uniquely in Australia, Victoria has maintained a collection of the end of year teacher judgment data for benchmarking purposes up to at least 2009. The case studies described here provide examples of the utility of teacher judgement assessments and some insights into the comparability of teacher assessments with test assessments.

Case 1- National Assessment –Language and Literacy

Masters and Forster (1997) used teacher judgement as part of an Australian national survey of students in years 3 and 5. The purpose of the survey was to establish an understanding of national English language skills in reading, writing, speaking, listening and viewing. The assessment methodology was unique in the way it linked classroom assessment into a national data collection process. Teacher judgment assessment, however, meant that this methodology was more costly than assessment processes dependent on external marking.[17]

Teacher judgment of student achievement was found to be reliable when supported by good assessment materials, professional development of teachers and the provision of advice from trained external assessors. Sample checking of teacher assessments, using two panels of markers (project staff and a team of trained markers) reviewed teacher assessments. The percentages of unchanged results were 98% for reading, and between 93% and 90% for viewing and listening. Changes were never more than one level. Correlations were calculated for between project staff assessment (in the range 0.91 to 0.99 depending on the task being assessed) and between project staff and teachers assessments (mostly above 0.8 with many above 0.9).

This project showed teacher-judgement assessment to be reliable, assumed to be of greater validity than paper and pencil testing since the characteristics assessed were observed over extended periods and of a quality more than adequate for broad system descriptions.

Cases 2 to 4- Victoria -the use of profiles for teacher judgement assessments

In Victoria teacher and test assessments have run concurrently using a common, regularly updated curriculum framework with a level structure. The arrangements have some close parallels with the England Key Stage assessments although commentators point out that the Key Stages are not vertically equated and thus limited in illustrating developmental growth (Masters, Rowley, Ainley, & Khoo, 2008). Tests have been conducted in Victoria at Years 3, 5, 7, and 9 in English/Literacy and Number/Mathematics starting in the mid 1990s with Years 3 and 5. Victorian developed tests have been replaced by national tests since 2008.

Over the same period departmental policy required teachers to record student learning status in the form of levels, at least once a year. These summative teacher assessments of students have been collected centrally at all year levels from Prep to Year 10 and have been reported

[17] For a survey where the costs include training, moderation, multiple visits to a site and cross-checking the cost is greater that a pencil and paper test. Were teachers already calibrated to a scale, with moderation already built in as part of the normal classroom processes, reporting of student learning status by teachers should be less costly than pencil and paper tests.

back to schools as a form of benchmark support to self-reference school data with state norms. This makes the Victorian department's data unique in the world. Compared with the England collections at three Key Stages, the Victorian Department has data at 11 Year levels per annum for more than a decade. There is currently no public process to compare teacher and tests assessments at an individual student level. However, with the introduction of the Victorian Student Number (VSN), a unique student identifier, in Victorian Government schools in mid 2009 (Victorian Auditor-General, 2009, p. 11) the matching of teacher and test data at the individual student level should be feasible.

Two case studies and a developing online assessment project have been identified in Victoria to illustrate how teacher judgement assessments have been applied. Teacher judgement data and test data are not easily found in the public domain. An overview of the data for the state provided in a recent Auditor-General's report (Victorian Auditor-General, 2009) is drawn upon, along with statistical reports. These data are described below. In addition the Quality Schools Project (Rowe & Hill, 1996) using teacher assessments as a prime source of data for monitoring learning changes over time is reported.

Case 2- Data from the Victorian Auditor-General 2009

The Auditor-General (Victorian Auditor-General, 2009) used the annual teacher judgement data and the complementary test data at Years 3, 5, 7 and 9 to report trends from 1998 onwards as part of an audit of Literacy and Numeracy programs. Data are presented in graphical form only without supporting tables. These graphs indicate the trends for teacher assessments and tests assessments over the period for reading and mathematics.

The two summative assessments are made at different times of the year (teachers in late Term 4 of 4, tests in early Term 3 up to 2007) and are collected through different processes. Up to 2007 Achievement Improvement Monitor (AIM) Tests were taken by all eligible students in the appropriate year levels, centrally marked and reported to students on the levels scale at scale divisions of 0.1 of a level. AIM tests have been replaced by National Assessment Program – Literacy and Numeracy (NAPLAN) tests and, as a result, are reported in 2008 and 2009 on a scale with no direct link to the levels structure.

Teacher assessments were collected electronically from school administrative records. Up to 2006, each strand-level was divided into three subdivisions (beginning, consolidating and established). Since 2006, coincident with the introduction of the *Victorian Essential Learning Standards (*VELS), the levels have been divided into four numerical subdivisions and reported as 1.0, 1.25, 1.5, 1.75 etc.. The time difference between when assessments occur, the lower scale resolution for teachers relative to the test scale and the change in the number of categories on the teacher scale make direct comparisons of the two data sources more

complicated. Appendix 4 to this thesis considers the likely effect of the increase in teacher response categories (from 3 to 4) and concludes that the change would be sufficient to generate lower state means, relative to the earlier period where three categories only were used, if no adjustment has been made to the time series. Based on the dips reported for 2006 and 2007 it is assumed no adjustment was made in the released data.

Teacher and test data in Victoria are reported up to 2007 on (notionally) the same scales, overcoming one of the issues with the US case studies. Trends over time can be compared, although the two processes for estimating students' positions on the scales, and quite different methods of collation, mean that it is unlikely that the mean scores for both processes would coincide exactly. The different collection times are dealt with by the Auditor General by adding 0.25 of a level to the test means, equivalent to half a year's growth (Victorian Auditor-General, 2009, fn. p. 75). This adjustment is compensation for the additional learning progress achieved by the time teachers' assessments are recorded. The time shift also highlights the issue of the degree of independence of the teacher assessment. The teacher has access to the student's test results in Years 3, 5, 7 and 9 before the final teacher assessment for the year is reported. Figure 4.3 compares teacher and test state means estimated from the Auditor-General's report (pp. 72 - 74).

Figure 4.3 Panel 1 presents the original teacher judgement assessment data extracted from the report with the points estimated from graphs C4 and C9. The final two teacher estimates (2006, 2007) are adjusted upwards by 0.1 of a scale position (based on a broad estimate of the likely effect of increasing the scale categories from 3 to 4 for each level –see Appendix 4) to create a second series of data points (Year 3 Teacher-2006, 2007 adjusted). The lines connecting the points and the regression lines are added to help reveal the trends. The adjusted points fit the general trend of the previous eight years but retain the downturn in the mean of the assessments indicated for 2007. Without the adjustment a marked downward shift is shown, inconsistent with the much smoother trend shown in the test data in Panel 2.

Panel 2 illustrates the adjustment for the time difference for the test. Raising the line by 0.25 of a level results in a close correspondence with the teacher data from Panel 1, charted together in Panel 3. The test data means show much greater amplitude of variation with time than do the teacher means. This greater consistency of teacher judgement is consistent with the data for England. Means of teacher assessments follow an upward trend up to the teacher scale category changes in 2006. The regression line for the original teacher assessments has a negative gradient in Panel 1, due to the effect of the last two points, and also a low R^2. When the last two data points (2006, 2007) are adjusted upwards by 0.1 of a level (based on Appendix 4), the gradient becomes positive and tracks in parallel with the test OLS regression

over time. The teacher gradient with calendar year is comparable to the test gradient. The R^2 value is also improved.

The regression lines for teachers and adjusted tests are almost parallel and only differ by about 0.02 of a level on average for any year. As shown in Panel 4 only a very small additional increase in adjustment to the original test data for the effect of different collection times would be needed (0.27 of a level rather than 0.25 of a level), to have the two regression lines effectively coalesce over the period from 1999 to 2005; and for 2006 and 2007 if the Appendix 4 scale adjustment is accepted.

No indication is given of the measurement error in the two processes. The estimation process for the points applied by the author, based on converting graphical plots back to estimations of the plotted values, has potential for adding further error. Assuming the combined error effects were as low as 0.075 of a level, the two data sets would be within the 95% confidence boundaries of each other for most of the pairs of points. On the basis of this tentative analysis, the trajectories of the estimates of the average teacher assessments and the test assessments for the same student populations are very close on the test scale, showing a gradient of improvement of 0.004 of a level per annum from both the teachers' and test perspective. Without presuming that the assessments for each individual student would be as close, it seems feasible to describe the overall trends in learning using either data source. The teacher assessments for a decade for Year 3 are very similar to the test assessments.

The plots for Years 5 and 7 exhibit the same general relationship, illustrated in Figure 4.4, once an adjustment for the revised scale categories for the teacher assessments for 2006 and 2007 is made. Year 3 means from both assessment sources are very close. At Year 5 teachers are assessing students consistently at about 0.12 of a level above the test. At year 7 teachers are assessing students consistently at about 0.07 of a level above the test. These are indications that teachers are not calibrated exactly to the test scale but, based on the consistency of the results, are remarkably close.

Figure 4.3 Comparison of Times series of Year 3 Teacher and Tests Data (values estimated from original graphs in Victorian Auditor-General, 2009)

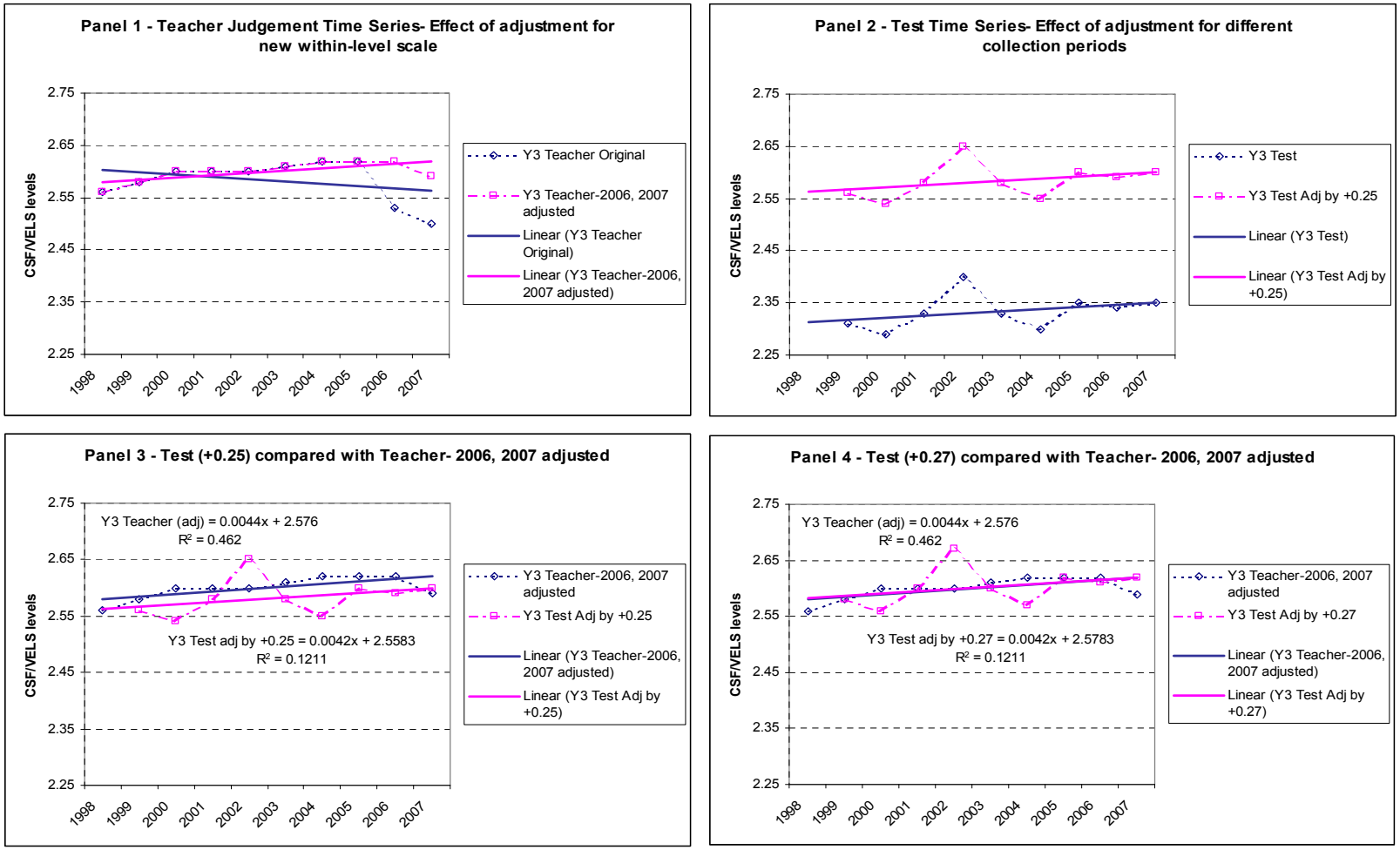
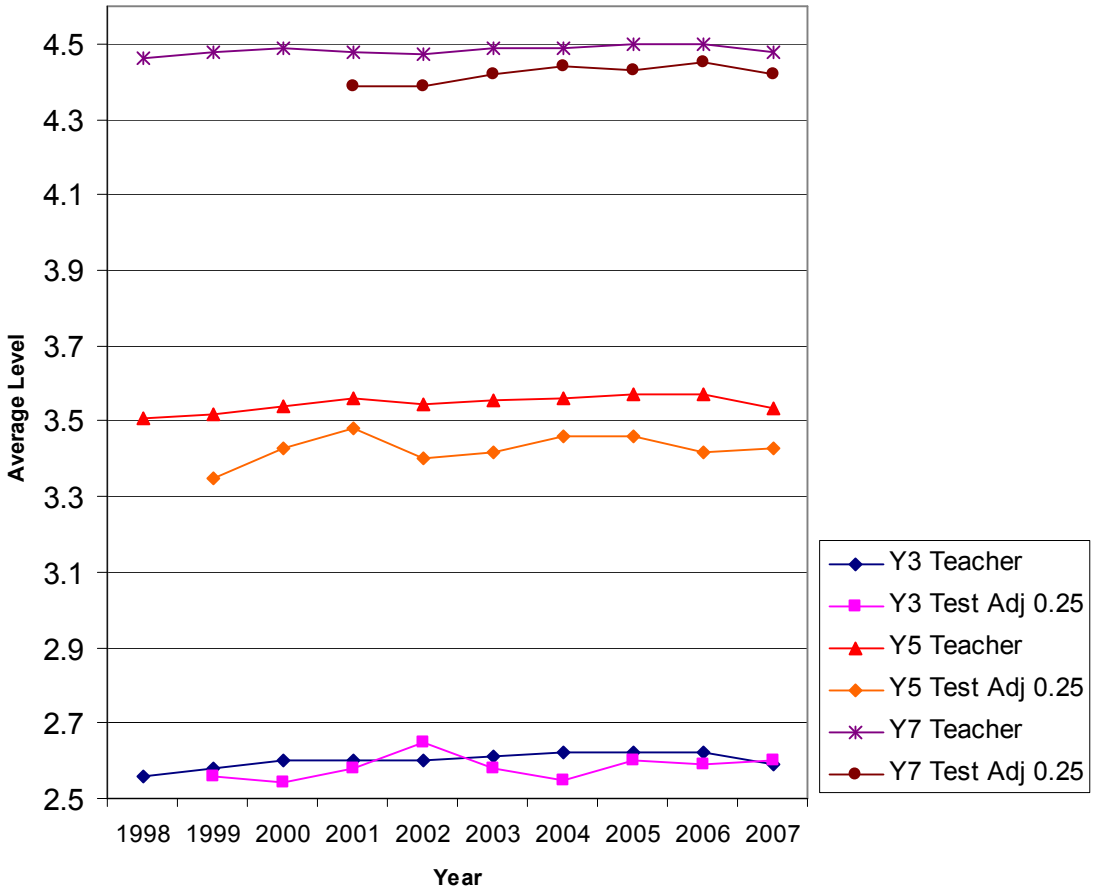


Figure 4.4 Comparison of Times series of Year 3, 5 and 7 Teacher and Test Data –Reading (values estimated from original graphs in Victorian Auditor-General, 2009)



Note: Teacher assessment grand means for 2006 and 2007 adjusted for the effect of re-categorisation of within level progress. Adjustment raises last 2 teacher assessments points by 0.1 of a level.

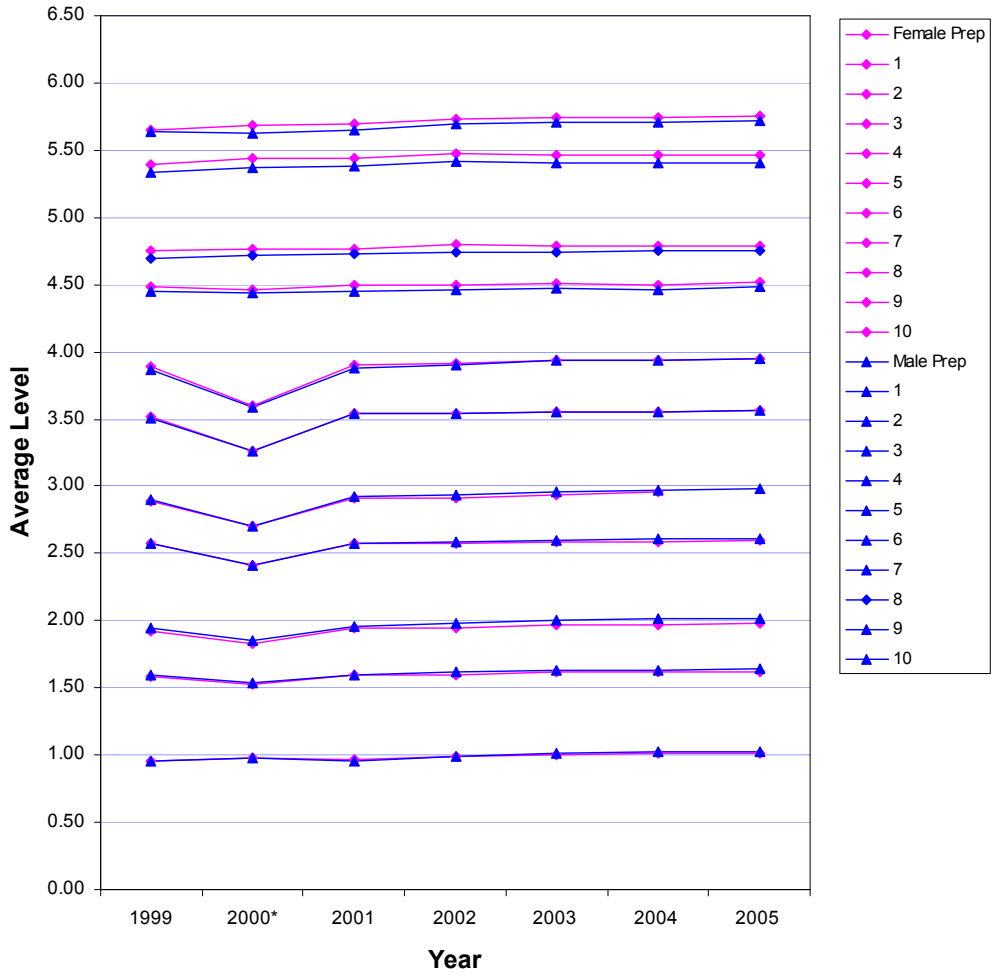
The two assessment processes, allowing for the general errors of measurement and estimation, produce very similar results at a population level. Where there are differences between the two sources of assessment, the differences are very consistent suggesting that there are possibilities for recalibrating either teachers or the tests. Based on the consistency and stability of the mean teachers' assessments it is possible that it is the test that should be re-calibrated. As raised earlier, the different scale categories (4 per level for teachers versus at least 10 per level for tests) will influence the error of the mean in each case. The England data illustrate that the degree of exact match of teacher and test assessments for individual students is likely to be only moderate, with most mismatches within a close range to the test assessments. The Victorian data for over a decade show that, overall, teacher and test assessments do match closely. With the introduction of unique student identifiers in Victorian Government schools in mid 2009 (Victorian Auditor-General, 2009, p. 11), the comparing of teacher and test assessments at the individual student level should be feasible, provided the test assessment can be rescaled to the VELS scale.

The Auditor-General had concerns about the broadness the teacher assessment units arguing that "progress that is assessed through teacher judgments could be improved, for example by increasing the number of progression points against which the judgments are reported" (Victorian Auditor-General, 2009, p. 61). How many progress points - that is how finely teachers can discriminate changes in student learning - is a topic needing further research. In principle it should be feasible for teachers and tests to have at least similar refinements in sensitivity for identifying increments of learning.

The Victorian school system has an extended time-series of teacher judgement assessments. The Auditor-General's report is one source for the data. Prior to 2006 the Victorian department released the data to schools through a website for reference purposes. Based on the contents of a small number of 2008 school annual reports (Caroline Springs College, 2008; Marist-Sion College, 2008), schools have continued to receive state benchmarks since 2006 but through a less public process. From the documents published in the period up to 2006 it is possible to build up a times series for English (reading) and mathematics by Year level. This sequential Year level view of mean learning status from Prep to Year 10 is unique in the world as far as this author can determine. Most systems in the US under the *No Child Left Behind* requirements have built cross-sectional data views but from Year levels 3 to 8 only, using test data. None report data from the commencement of school through to Year 10, possibly due to the cost of collecting the data and the inappropriateness of pencil and paper tests at Prep, Year 1 and Year2 (K, Grade 1 and Grade 2 in US terminology).

The very regular relationships of Year levels to each other since 1999 are shown in Figure 4.5 for mathematics and in Figures 4.6 and 4.7 for reading (Department of Education and Early Childhood Development, Victoria, 2003, 2006). This view of the data reveals the consistency of the mean results over this extended period, and across learning areas. The sole aberration in the general pattern is the one-year effect demonstrated in 2000 in mathematics (Fig. 4.5). In this year the original Curriculum Standards Framework (CSF1) was replaced by CSF11. In the change the convention for reporting mathematics in P-6 was altered. This led to apparently aberrant means for 2000 at Years 1 to 6. Once reporting adjustments were made during 2001, the series for Year levels 1 to 6 resumed trends very consistent with the position in 1999. The companion time-series for English shows the same Year level relationships without the deviation for 2000 (Fig. 4.6).

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



Source: Department of Education and Early Childhood Development, Victoria, 2003, 2006

The mean CSF levels reported for each Year level have remained consistent with a small growth trend from 1999 to 2005, well illustrated later in Figure 4.7 for reading. The average growth over all Year levels from 1999 to 2005 is approximately 0.05 of a level (0.008 per annum), though there are variations by Year level. Time series longitudinal patterns (diagonal change of the same students) and cross-sectional patterns (horizontal change in cohorts from one year to the next) are very similar. From Figure 4.5 the spacing between the lines shows the cross-sectional growth. (Figure 4.6 shows that view of reading data in a slightly different form of presentation). A consistent pattern applies in Figure 4.5. Prep to Year 1 growth is relatively large, about 0.6 of a level. Growth from Year 1 to 2 is about 0.4 of a level. This alternating pattern of less growth in particular periods (1 to 2) and then more growth in the next period (2 to 3) is maintained over the P to Year 10 spectrum and over all calendar years (ignoring the 2000 aberration). The same pattern is shown in Figure 4.6 by the consistent placement of data points above or below the regression line for 1999 as a reference line. Implied is that teacher judgement assessments (in the Year levels where tests apply; 3,

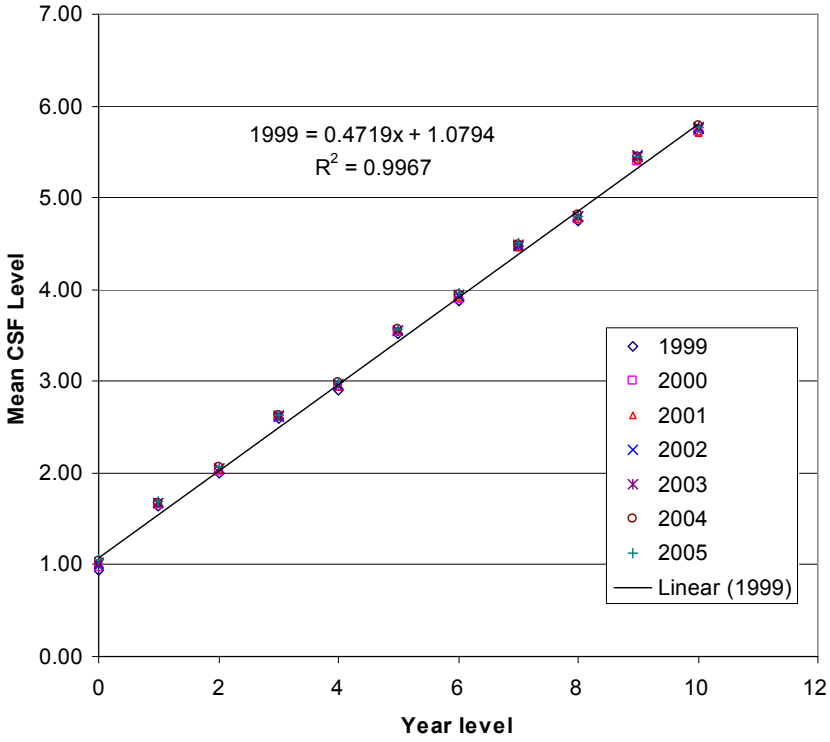
5, 7 & 9) are slightly higher on average than where they would be if the annual growth per Year level were even. The values in the tested Year levels may be affected by the moderation effects of the feedback of test results to teachers in these Year levels.

Without the provision of regular and consistent teacher judgement assessment data the pattern of learning across Year levels would not be appreciated. The general pattern has implications for understanding learning growth and so it is important to consider the possibility that the pattern is an artefact of the collection process or some other systematic error. The data are a summary of approximately 440,000 independent teacher assessments per annum through collection processes that have changed over time. The likelihood of a systematic collection error is very low. A comparison standardised assessment process at all Year levels (i.e. a test) would be one method to confirm the inter Year level patterns. Even with a sample approach this would be a large undertaking. The pattern might also reflect subtle variations in the calibration of teachers for given Year levels. The consistency of the linear growth by Year level over many years is quite a remarkable phenomenon reinforcing the regularity of teacher assessment overall and the value of actually having such data.

The trends indicate slightly higher scores for females in Years 7 to 10 in mathematics. Nationally reported test data for Victoria for 2008 and previous years (National Assessment Program Literacy and Numeracy, 2008; National Report on Schooling in Australia Preliminary Paper, 2007), shows the reverse position by gender for Year levels 7 and 9 in Victoria. The teacher-reported higher scores for female students suggest a small teacher assessment bias in perceiving the performance of girls. The effect is small but has been persistent over time. It is beyond the scope of this thesis to investigate whether tests have an opposite bias or the event of test taking has a differential impact on girls at higher year levels, impeding the expression of their most likely learning status position. In further consideration of teacher judgement assessments, resolving which of the possibilities applies is necessary for understanding a subtle but important validity issue. More importantly resolving whether there is a gender bias, or other influences from SES, language background and so on where systematic differentiation may be noticed in either assessment process will impact the training and calibration of teachers, if teacher judgement assessment can be shown to be a feasible source of longitudinal assessment data.

A similar general pattern in the same Year levels applies in the equivalent English Learning Area (reading) series by gender (not shown here -see Figure 4.6 for the all students pattern). Female students in English achieve a slightly higher average score at all year levels, with the gap increasing from P to Year 10. In this case the pattern is consistent with independent test data (National Assessment Program Literacy and Numeracy, 2008), in which girls score a higher average score at all Year levels.

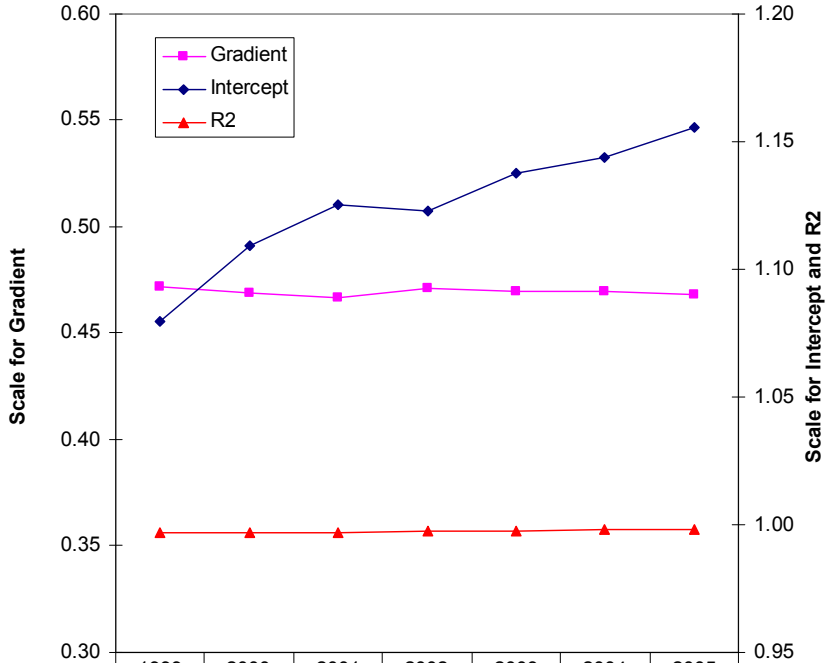
Figure 4.6 Reading: All students-Mean Teacher Judgement Assessments 1999-2005 by Year level.



The consistency of teacher assessments over time and Year level is illustrated in Figures 4.6 and 4.7. Figure 4.6 shows the nearly linear relationship of mean teacher assessed CSF level with Year level, consistent from 1999 to 2005. The one regression line (Years P to 10) for 1999 is shown as a reference.

Figure 4.7 shows the trends in intercept, gradient and variance explained (R^2) calculated for the regression from P to Year 10 for each of the calendar years over the period 1999 to 2005, represented in Figure 4.6. The consistency of the gradient and intercept for the OLS regression by Year level over seven annual replications is very high. The parameter that has changed the most is the intercept; it has moved consistently upwards, implying a general teacher perceived improvement in student learning status at lower Year levels over time. There is a slight decline in the gradient of the growth rate with Year level over the same period implying that the rate of learning increase is slightly less in the upper Year levels. The fit of the line through the means (as represented by R^2) has remained consistently high.

Figure 4.7 Reading: All students-Plot of regression parameters for each year 1999 to 2005.



	1999	2000	2001	2002	2003	2004	2005
Gradient	0.47	0.47	0.47	0.47	0.47	0.47	0.47
Intercept	1.08	1.11	1.13	1.12	1.14	1.14	1.16
R2	1.00	1.00	1.00	1.00	1.00	1.00	1.00

These Victorian data illustrate a consistency in teacher reported data over time but also indicate a small and relatively smooth improvement in mean learning status in reading and mathematics (and writing and speaking and listening). Compared to test data over the same period (Figures 4.3.and 4.4), the variations in the grand means of the test in successive collection periods, relative to the general trend, are greater than those shown in the teacher data. The larger units used in teacher judgement assessments may account for this. The trajectory of the teacher means is generally smoother and similar in this feature to the trends for England illustrated in Figure 4.1. While the grand means of teacher and test estimates of learning do not coincide exactly, the general patterns and trajectories are similar. The stability of the patterns implies that teacher assessments are consistent indicators of something. Differences by gender for teacher assessments are consistent with the test patterns in English/literacy, although the overall combined means for tests and teacher assessments are displaced as illustrated in Figure 4.4. There is an indication of a possible bias in higher Year levels in numeracy/mathematics in favour of girls when teacher judgment assessments apply.

Case 3- Quality Schools Project

Teacher judgment assessments as the measure of learning improvement were used in the Victorian Quality Schools Project (VQSP). Rowe and Hill (1996) used a profile approach (as described in Chapter 3 and above for Victorian schools) as the basis for monitoring student development over a number of Year levels. Teacher judgements of student learning development were used as the dependent variable for monitoring learning. The VQSP longitudinal study obtained data on educational progress in English and mathematics for entire year-level cohorts from Kindergarten through to Year 11. A sample of 13,900 students, drawn from 90 government, Catholic and independent primary and secondary schools, were the subjects of the study.

Comparisons of the teacher-mediated assessments with independent assessments (say by appropriate tests) were not part of the investigation. The researchers applied the Guttman process for estimating true reliability and obtained values ranging from 0.67 to 0.81 in Reception (Kindergarten) to 0.9 to 0.92 at Year 11. The results indicated that the profile strands appeared to function as cumulative scales or growth continua and that teachers were consistent in their use of them. Test/re-test reliability estimates were made using correlations (Pearson's r) between teacher assessments for the same students made four months earlier. The correlations indicated that teachers assessed their students consistently when asked to provide a repeat assessment. Values ranged from 0.89 in Year 1 through to 0.92 in Year 11. Limited evidence regarding inter-rater reliability was provided when two or more teachers serendipitously rated the same student. Inter-rater correlations ranged from 0.85 to 0.89. At the level of precision required by the profiles, teachers were regarded as consistent assessors.

The data enabled, among other descriptions and analyses, elegant graphical presentations of the progress of students through the Year levels (similar to Figure 4.5), along with the spread of the development at any Year level[18]. At any point the progress scale can be interpreted into what students at this point can do.

The study illustrates that teachers, with appropriate frameworks, are able to estimate student reading and mathematics learning developmental status. The scale underlying the assessments was treated as an interval scale, with the VQSP Band descriptions being regarded as vertically scaled. The descriptions of growth across Year levels and the spread of growth within Year levels were developed cost effectively without the use of tests.

---

[18] Comparison with Rowe and Hill (1996, p. 332) shows Year level to Year level growth patterns do not match those shown in Figure 4.5 at the specific year levels. They do however show a general alternating pattern of growth for consecutive Year levels and then less growth for the next Year level.

Another Victorian evaluation study used a similar approach to that of Rowe and Hill. The Literacy Advance In The Early And Middle Primary Years Project (Ainley, Fleming, & McGregor, 2002) used teacher judgement assessment and confirmed high correlations between teacher assessments and those of external trained assessors. An earlier Victorian study (Sharpley & Edgar, 1986) is regularly cited. This predated the use of levels and compared teacher ratings of students with standardised tests; the Progressive Assessment Tests (PAT) and Peabody Picture Vocabulary Test-Revised (PPVT-R). Correlations between teacher ratings on a five point scale and test scores were generally in the range from 0.4 to 0.5. A possible bias in favour of girls in teacher assessments was reported.

Case 4- Online structured assessment interviews that require teacher judgement

The processes used by Victorian teachers to estimate learning status have been dependent upon their interpretation of the general frameworks (CSF 1, CSF 11, VELS). The results described above indicate that teachers, as a group, appear to be consistent in their judgements, assuming stability of overall judgments and parallel trends to test assessment as reliability and validity criteria.

New assessment tools have been developed that among other functions can help maintain this consistency. One approach to improving consistency has been to provide teachers with online interview protocols that lead to estimates of the learning status of a student. However, there is a risk that these tools could become too specific and time consuming and negate the general *connoisseur* element of informed expert judgement.

From 2007 to 2009 online interview guides in English for Prep to Year 2 (Department of Education and Early Childhood Development, Victoria, 2009a) and mathematics (Department of Education and Early Childhood Development, Victoria, 2009b) have become available. The English interview is compulsory for all students in the early years. Students are assessed at the start of Prep, end of Prep, end of Year 1 and end of Year 2. The teacher uses a web-supported process to interview students and record their skill levels on a number of aspects. Among ensuing reports is a longitudinal report for each student, reported on the VELS levels scale, with the learning status estimate being in 0.1 divisions of a level. This is further evidence that smaller subdivisions for the teacher level scales are required and are likely to be practical.

The links to the VELS scales seem to be maintained in the face of environmental changes such as the introduction of the NAPLAN tests on different scales. Conversion tables for NAPLAN scores are provided (based examples from school annual reports and Student

Performance Analyser-SPA-software[19]) so that NAPLAN test scores for some strands can be converted to the VELS scale, maintaining the link across assessments for schools for their longitudinal data. A consequence of the impending national curriculum (National Curriculum Board, 2009) may be a structure description that does not include levels in the form used in the VELS. As a result, maintaining assessment records and simple teacher judgement assessments across Year levels and calendar years may be disrupted.

Over the set of Victorian examples cited above, the use of the test scales (calibrated in VELS levels in the most recent examples) has proved feasible as a method of recording summative assessments of the learning development of students. Data points are separated by 4 to 12 months when teacher and test assessments are viewed together. Test assessments are 2 years apart for individual students. The means of the teacher assessments are shown to be very similar to test assessments, under particular assumptions to bring them to common time points and remarkably regular in the trends on an annual basis. The general consistency suggests that a system of recording student learning based on teacher judgement assessments using the test scale might be feasible.

Case 5-Tasmania -Indirect evidence- validation studies

One Tasmanian source provides an indirect insight into the quality of teacher judgement assessments. Callingham (2003) investigated the validity of a performance measure, assessed directly by teachers. Findings indicated that the performance assessment validity was high. Students in Year 10 from 14 government high schools in Tasmania undertook an assessment battery that included a performance task assessed directly by teachers using a rubric, a multiple choice test of mathematics skills, and an objective test of mathematical problem solving. Teachers were also asked to rate their students' mathematics ability on a Likert scale instrument with 10 items, to provide an additional teacher judgment measure. This approach therefore included the direct (performance scale) and indirect (Likert scale ratings) concepts of Hoge and Coladarci (1989).

The assessments explored two different traits. The performance task and the problem-solving test addressed the same trait, higher order thinking. The skills tests and the teacher rating of mathematics ability addressed a second trait, mathematics ability.

The teacher rating of mathematical ability showed considerable underfit to the Rasch model used in the analysis, for some teachers. This was interpreted to mean that there were a number of students for those teachers where the teachers' judgments of students' mathematics abilities were erratic ("affected by randomness" according to Callingham, p. 14). However, it

[19] See examples from SPA website http://www.sreams.com.au/home_page.html.

> … appeared that teachers made similar overall judgments about their students' ability to that determined through the performance assessment task but that these judgments were more consistent when made against a scoring rubric, rather than made as an holistic judgment on a rating scale. (Callingham, 2003, p. 14)

This is consistent with the findings of the benefits of direct assessment (Hoge & Coladarci, 1989). There are indications from inspection of Callingham's graphs that a small number of teachers vary in their judgments, when compared with the test assessments but that overall most teachers compare well with the tests. Correlations between assessment methods are in the range of 0.45 to 0.57, except for the higher-order thinking test versus the mathematics ability test where the correlation is a higher 0.78.

All students in a class were assessed under all four methods (c.f. the small samples used in many of the US cases cited earlier). Thus the data provide an opportunity for an understanding of the degree of match of methods at the individual teacher level using all students for each teacher. All assessments for all students were estimated in Rasch logits. On this basis it should be feasible to compare teacher assessments with test assessments using the 45-degree line method to establish the assessment accuracy of individual teachers and the extent to which individual teachers are matched to the test scale. This was not the purpose of the original study. However the data from the study have the rare potential to identify and quantify the degree of spread in the ability of teachers to match their judgements to the test judgements, through the analysis of each teacher's full class data. The study, overall, confirms that teacher judgement is a valid assessment process and that a direct assessment (rubric) is more consistent than an indirect teacher assessment (rating scale).

Cases 6 & 7-Queensland and the Australian Capital Territory

Teacher judgement assessments of students apply informally in all Australian school systems, particularly in the Years K to 10. Two school systems use mostly teacher assessments for Year 12 certification, rather than subject-based external examinations. These systems, Queensland and the Australian Capital Territory, are seen as world-leading examples (Harlen, 2005a). Other Australian systems have a mix of school based and externally examined subjects at Year 12 (Victoria, South Australia and Northern Territory are examples).

Queensland's Year 12 courses are based on subject syllabuses, broad frameworks that allow flexibility of local implementation. Each subject is developed into a teaching and assessment plan (work plan) by the school. The criteria and standards matrix for final (exit or end of course) levels of achievement for recording on the Year 12 Certificate are stated in the specifics of the school's work plan. Assessment processes are designed by the teachers to be appropriate to the intended learning outcomes.

> All subjects involve other forms of learning than those assessable through written examinations. This expansion of the forms of learning (and thus forms of assessment) was one of the original intentions of the move to school-based assessment. It allows for more authentic assessment to occur, connecting with student interests and making learning and assessment more meaningful and applicable for students. (Maxwell, 2004, p. 2)

Assessment in the ACT is also school based. No examinations are set by a central authority for any subject but, as in Queensland, there is a generic skills test (*ACT Scaling Test*, AST) to measure skills deemed necessary for success at university (*ACT Board of Senior Secondary Studies Policy and Procedures Manual*, 2009). In Years 11 and 12, courses are taught and assessment is conducted and recorded unit by unit.

The tests applied in each of the two systems (Qld. and ACT) are generic and not intended to directly validate the teachers' assessments. They do not relate to any particular subject teacher's view of a student's progress. Both systems have used teacher judgement for a long period and would appear satisfied with the quality assurance and moderation processes applied to establish comparability of teacher assessments. In both cases these comprehensive school and system moderation arrangements add to the capacity of teachers' making on balance assessments. Neither system has collections of teacher judgement assessments at primary level of the form considered later in this thesis.

At the primary school level in Queensland, Cumming, Wyatt-Smith, Elkins and Neville (2006) investigated teachers' assessment practices in literacy and numeracy in Years 3 to 6. They interviewed teachers in seven schools, focussing on 70 students. The purpose of the investigation was, among others, to consider the extent to which the outcomes of Year 3 and 5 tests and teacher judgement assessments were congruent or differed. Ways in which the two assessment processes could be used as complementary sources of data, to support both improved learning outcomes for students and systemic data collection were considered. The focus of interest is very similar to that of this thesis.

Teachers considered "similar dimensions of literacy and numeracy to those measured by tests, although it is not possible to determine whether these judgments and the teacher discussions were influenced by the project focus" (Cumming et al., 2006. p. 7). Both teacher assessments and tests were regarded as narrow in their skill attention, in comparison to the broader policy and official curriculum frameworks for literacy and numeracy. Teacher judgments of student levels were found to be broadly consistent with outcomes for individual students on the Year 3 and Year 5 tests. Teachers also indicated that where there were divergences, teachers considered their own judgments to have a more substantial basis.

The researchers noted the lack of preparedness of the seven schools to track individuals and cohorts systematically using data over time for longitudinal analyses of student performance - another concern of this thesis. They saw a need for school leaders and system personnel to develop their ability to support schools to improve the use of existing system data. In particular, school leaders and system personnel need to support schools to optimise the test data, and to examine its coherence with locally generated assessment information.

Overall the Queensland and ACT systems provide evidence that consistent summative assessments can be made by teacher judgement and that at the primary level in Queensland there is evidence that teachers and test assessments are broadly consistent.

Case 8- South Australia

Teacher judgement assessment is common practice in South Australia. A general history of the 1990s period is covered in Chapter 3. Formal collection of data on teacher judgements applied in 1997 and 1998 only. The data are presented in Chapter 7. Statewide testing of students has applied since 1996. No comparison of teacher and test assessments, apart from that in this thesis, has been made. However one small study, interested in the relationship of teacher judgements using the reading profiles from the Statements and Profiles for Australian Schools compared to a standardised reading test, was carried out in 1997.

Bates and Nettelbeck (2001) researched the same population of South Australian primary teachers in the same year (1997) as analysed in this thesis. Their study provides an example of the difficulty in bringing teacher judgements and test results to a common scale. Bates and Nettelbeck explored the ability of teachers to estimate reading achievement. The procedure required teachers to be aware of the assessment process and norm concepts of the *Neale Analysis of Reading Ability-Revised* (NARA-R). Bates and Nettelbeck report that the NARA-R is an instrument with which, it would seem, the teachers had not had previous contact. Teachers were

> provided with written information about the NARA-R, including scoring procedures and relationships, set out in a table, between raw scores, reading ages and age-corrected percentile ranks. Instructions emphasised the concept of percentile position (for example, 'this child achieves at a level better than almost 75% of children'; or 'about 60% of children outscore this child'). (Bates & Nettelbeck, 2001, p.180)

Bates and Nettelbeck go onto explain, "all teachers confirmed that they understood what was required" (p.180). Teachers were then asked to estimate the percentile ranking of each student in the sample for the teacher (3 or 4 per teacher). This estimate appears to have been in terms of the national norms, not just where in the class or where in the school but, from their (assumed meagre) understanding of the NARA-R national norms, the percentile

placement of each student on the national norms. Perry and Meisels (1996) indicate that this form of norm judgement is not easy for teachers.

There was no explicit reference frame for teachers who had to make judgements solely in terms of the national percentile norms. This percentile was then reverse interpreted to generate a raw score, and then a reading age. This would appear to be a very complicated (or at least fraught) process from the teacher's perspective, though the difficulty in getting the teacher judgement and the test score onto the same scale is also appreciated.

One of the researchers then independently assessed the students with the NARA-R to generate the test data. That there were mismatches in the range of 18 or so months either side of the test-established reading age is not overly surprising, given the possible errors in the assessment procedure. With respect to reading accuracy the matches were spread as follows: above or below by 0 to 5 months –23%, 6-8 months-26%, 9-12 months-18% and greater than 12 months -32%. An approximately similar pattern applied for comprehension.

On these data, teachers were within 8 months of the test assessment (within 1/3 of a SPFAS level) for approximately 50% of cases. On the assumption that the 32% of cases more than 12 months away from the test result were distributed as the tails of a normal distribution, an estimated 10 to 15% of cases were more than 2/3rds of a level away from the test assessment. This degree of match is less than Table 4.3 (Durant, 2003) for England (88% within 1/3 of a level in reading versus 50%). However based on the concept of levels as it applied in SA at the time (no subdivisions within a level), had the teachers been asked to assign levels it is likely that about 65% of assessments would have been in the same level (i.e. within 12 months, above or below, the test assessment).

The complicated scoring method is most likely to have been a source of error for the teacher assessments. Bates and Nettelbeck report that teachers' estimates of percentile placement in reading were moderately correlated with Neale reading accuracy (0.77) and reading comprehension (0.62) test scores, consistent with the range found in other studies summarized by Hoge and Coladarci (1989). What the data do not provide are indications of the relationship of the order of the teacher judgements for each teacher for their sample of students, and whether the spacing of assessments bears any relationship to the original reading age scale (Criterion 2 above).

Among other findings, the researchers concluded that teachers tended to "over-estimate the relative percentile position of children performing less well and under-estimate the achievement of better readers" (p. 183). As argued above, the judgement or estimation process was dependent upon the teacher already having a reading age reference frame (or more distantly a percentile reference frame), the same as estimated by the independently

applied test process by the researcher. This requirement for teachers and tests to be in the same reference frame makes the judgement task problematic when the reference frame is unfamiliar or not well understood. The researchers concluded that there

> would appear to be a need to implement a structure that explicitly sets out the standards that students are expected to achieve. Logically, teachers cannot be held accountable for students' performance levels if they are not provided with the information necessary to uphold such ideals. (Bates & Nettelbeck, 2001, p 185)

This is a laudable conclusion and certainly reflects other critic's views about the need for added clarity in the description of learning criteria within the framework of the profiles in South Australia. Whether the comparison of teacher judgements with the NARA-R was a fair test of teachers' judgements is another matter, particularly important now since the Bates and Nettelbeck paper has recently been cited by a number of other researchers (Canto, 2006; Laidra, Allik, Harro, Merenäkk, & Harro; 2006; Freund, Holling & Preckel, 2007; Gilmore & Vance, 2007; Triga, 2004). On the basis of the use of an unfamiliar instrument and scale and the complex process to derive a score value, the Bates and Nettelbeck evaluation of teacher judgement skill might overstate apparent inadequacies.

Do accurate teacher assessments influence learning?

A major reason for a considering the ability of teachers to make assessment judgements that match those of tests is the claim that this skill might influence student learning. Hardly any studies, as far as could be determined, explore teacher assessment accuracy in this exact paradigm and its effect on learning.

Literature on the value of formative assessment and the cost effective benefits to student learning it can provide, are covered widely (Bangert-Drowns, Kulik, Kulik & Morgan, 1991; Black & Wiliam, 1998a, 1998b; Brookhart, 2004; Fuchs & Fuchs, 1986; Hattie & Timperley, 2007; Leahy & Wiliam, 2009; Shute, 2008). Formative assessment and feedback have been shown "to improve students' learning and enhance teachers' teaching to the extent that the learners are receptive and the feedback is on target (valid), objective, focused, and clear" (Shute, 2008, p. 182). The general finding across school subjects, countries, and ages is that "formative assessment appears to be associated with considerable improvements in the rate of learning" (Leahy & Wiliam, 2009, p. 3). Wiliam and Thompson (2007 cited in Leahy & Wiliam, 2009, p. 3) estimate that formative assessment is likely to be twenty times more cost-effective than programs that reduce class-sizes, suggesting that formative assessment is likely to be one of the most effective ways of increasing student achievement.

Techniques for assessment described in the assessment studies reviewed include teacher observation along with structured classroom processes for the teacher to be sensitive to the

understanding achieved by students. Short-cycle formative assessments, that is those conducted from two to five times per week can significantly improve student learning (Wiliam & Thompson, 2007; Yeh, 2006). Black (2003) on an even shorter time-scale using "in-the-moment" formative assessment (ongoing real time observations and probes by the teacher) found by all assessments applied that substantial gains in student achievement were achieved.

In the moment assessment comes close to the understanding of teacher judgement assessments used in the thesis, that is teachers holding a theory about how groups (and when feasible individuals) are progressing and having a scale to articulate and record this. None of the formative assessment reviews and studies draw on the concept of an underlying dimensional structure for learning progress nor a teacher's calibration to it. As a result there is little information on whether there is a benefit from the accuracy of teachers' judgement of where the students are located on such dimensions.

One recent study (Herman & Choi, 2008), explores the contribution of teacher accuracy in assessing the general progress of science classes and the teacher's accuracy in estimating the spread of the class across levels identified in a progress guide. A progress guide is described elsewhere in this thesis as a progress map or learning progression. Data from seven teachers were analysed, acknowledged by the researchers as rather small and "more of a case study rather than a firm empirical base" (p. 8), to offer some insights into the link of assessment accuracy to student learning. About 190 students were involved.

Herman and Choi concluded that teachers who estimated the general distribution of the learning status of their students most accurately demonstrated greater student learning growth through the unit. The effect was small. Improvements in accuracy of estimation of 10% increased the outcome score for students by up to 0.25 of a standard deviation. They established that some teachers were consistently better than others in their estimates but that all teachers showed inconsistency. There was considerable room for improvement in assessment accuracy, based on the differences between teachers assessments and those of the researchers. They assert that accuracy in assessment seems to be a necessary precursor to the use of assessment results in decisions about student learning.

The study supports the key ideas in this thesis. Knowing where a student is in his or her learning is a critical prior skill to being able to support that student's learning. The evidence in general from this chapter suggests that improvements in the ability of teachers to make formative assessments depends upon a deep knowledge of student learning patterns and that teachers differ in the extent to which they have this knowledge.

A summative overview of teacher judgement -Harlen

A summary of the extensive work of Harlen on the merits of teacher judgement provides an appropriate bookend to the chapter. Harlen has been involved in curriculum design and assessment over a number of years. She consolidated, in various documents, the research on teacher judgement. Harlen (2005b) summarises the findings of a systematic review of research on the reliability and validity of teacher assessment used for summative purposes. The original review (Harlen, 2004b), conducted under the auspices of the *Assessment and Learning Research Synthesis Group* of the Evidence for Policy and Practice Information and Coordinating Centre (EPPI-Centre), includes some of the studies mentioned earlier in this chapter (Coladarci, 1986; Meisels et al., 2001; Rowe & Hill, 1996). Prior to the teacher assessment review, Harlen (2004a) documented a wide range of research on the impact of assessment on students, teachers and the curriculum. The insights from these reviews are reported in detail in Harlen (2004a, 2004b, 2005a, 2005b) and used in forward-looking analyses in Harlen (2007a; 2007b).

Speaking specifically on the issue of teacher judgement assessments and whether they can be trusted, Harlen (2005b) resolves that, on balance, teacher's assessments can be trusted subject to specific support arrangements that include; identifying detailed criteria linked to learning goals, support for teachers' understanding of learning goals, professional development, moderation, time for planning assessments, and developing an assessment culture where assessment is seen positively and "not seen as a necessary chore" (Harlen, 2005b, pp. 267).

On the other hand she concludes that there is also "error and bias in teachers' judgements, … clearly revealed in some studies (Bennett et al., 1993; Brown et al., 1996, 1998)" (Harlen, 2005a, p. 265), which she claims can be addressed though training, and moderation of teachers' assessments. The referenced studies on bias concentrate on the upper levels of schooling, at the level of school completion and university entrance rather than on the beginning and middle stages of schooling. The extent of bias at these earlier levels is less clear, notwithstanding some competing arguments of bias (Rosenthal & Jacobson, 1968; Cooper & Tom, 1984).

The role of the teacher as a skilled professional is enhanced where teachers exercise their judgement in assessment. Harlen argues that

> using teachers' assessment [in summative assessments] gives teachers a genuinely professional role in assessment rather than one of merely following the directions of an external authority. Moreover, it means that teachers develop skills that will help them in gathering information that can be used for formative purposes, to help learning, as well as gathering information for summative assessment purposes. (Harlen, 2005b, p. 266)

Harlen sees the major risk in teacher judgement where the results of assessment are used for high stakes evaluations of teachers and schools. She argues that student assessments, whether by teachers or by external tests, are deficient as measures of teacher and school effectiveness. Other information is needed for understanding teacher and school effectiveness and this can be provided "without the damaging impact on students, teachers and the curriculum" that tests can make (Harlen, 2005b, p. 266).

In a commissioned report Harlen (2007a) consolidates her insights for an effective teacher based assessment process into a report for the Primary Review, an independent enquiry into the condition and future of primary education in England. She develops a critical review of the assessment system in England, describes how the various purposes and uses of assessment are met there, and in the other countries of the UK and in France, Sweden and New Zealand. Alternative methods of conducting student assessment for different purposes are considered in relation to their validity, reliability, impact on learning and teaching, and cost. She proposes the use of teachers' judgements as part of a future system design, as an alternative to depending on test results. She argues that since teachers can collect evidence during the numerous opportunities they have for "observing, questioning, listening to informal discussion and reviewing written work" (Assessment Reform Group-ARG, 2006, p. 9), this process at once

> not only improves validity but removes the source of unreliability that tests cannot avoid since they can include only a narrow sample of the learning goals. A particular advantage is that teachers will be gathering this information in any case if they are using assessment for learning. (Harlen, 2007a, p. 26)

In summary, Harlen (1994, 2004a, 2004b, 2005a, 2005b, 2007a, 2007b) provides arguments and evidence that support the building of systems of student learning on the judgements of teachers consistent with the thought experiment in this thesis. Teachers require a range of developmental supports to achieve this, including descriptions of levels of achievement, training in assessment and processes of moderation to ensure consistency of views on learning.

Summary

Evidence in this chapter indicates that teacher judgement assessment is an accepted and supported form of student assessment in a small number of jurisdictions. Cited studies confirm reasonable reliability in teacher judgement assessments of students as part of routine classroom activity. Teacher judgement assessment is used as part of formative assessment processes as well as in more formalised and standardised summative assessment activities. A

number of school systems have incorporated teacher judgement as a key part of their summative assessment at various year levels.

Two systems stand out as having the ability to compare teacher and test assessments. These are England using the Key Stages, and Victoria, Australia with the VELS (formerly CSF) assessments. Current teacher antagonism towards tests in England may mean that England's ability to compare results might be about to disappear there, and along with it, one option for moderating and maintaining teacher calibration. The new NAPLAN testing and National Curriculum in Australia may have a complementary effect in Victoria reducing the comparability of teacher judgement assessments with test results. The evidence from Victoria shows close mutual tracking of teacher and test assessments. Data by gender for teacher assessments in reading show the same trajectories as test assessments by calendar year and by Year level. There is also evidence of a small bias in teacher assessments in favour of girls in higher Year levels in the assessment of mathematics, relative to the test results.

In the England Key Stages teacher and test assessment also have approximately parallel trajectories by calendar year. This implies both assessment processes follow the same general trend. Mean teacher assessments however show less variability around the general trend lines. The relationships of teacher assessments to test assessment are subject dependent. In some subjects the teacher-assessed scores are consistently above the test score, in others the reverse occurs. Apparent systematic differences between the two assessment processes by subject leave unanswered the issue of which assessment process in each subject is likely to be the better estimate.

While calendar year tracking patterns of the averages of the data calculated independently for teachers and tests are close, more detailed analyses of the matches for individual students show a more moderate match rate, decreasing with higher Year levels. Whether the increasing mismatch rates are due to the inadequacy of test assessment or teacher assessment cannot be determined. Furthermore, given the lack of detailed analyses it is impossible to know whether teachers all mismatch at the same rate, or a smaller proportion of teachers account for a disproportionate share of the mismatches.

In all school systems where levels have provided the reference framework and scale for teacher judgement assessments, the assessments have been mainly summative. The finest resolution available in these scales is approximately 6 to 8 months of learning development. At this unit size the scales have little value in supporting formative assessment.

Whatever the quality of teacher judgement assessments, the reality is that teacher judgement is a large component of the educational process at the classroom level. A consequence of this reality is the requirement, as advocated by Harlen, to design schools, system and classroom

processes that capitalise on the primacy of this teacher judgement by ensuring adequate supporting criteria, adequate time for sharing of information about criteria and students and time to be coached in processes to fine tune judgements. This thesis argues as well the value of tests as support for teachers, to help in moderation and calibration.

The case studies described in the chapter indicate that a large number of teacher judgements, if made within supportive and well researched curriculum frameworks, can be comparable with test assessments. Only a small number of school systems have formalised the use of teacher judgements in reporting about student learning status and in these systems the two-way feedback to teachers from tests designers, and vice versa, does not appear to be established. Rowe and Hill argue that

> our … mistake as educationists has been the abrogation of our professional responsibility for the evaluation of students' educational progress by placing all our assessment 'eggs' in the psychometricians 'basket'. In so doing, we have devalued teacher training and professionalism, together with the experiences and rich contextual understandings that is their 'stock-in-trade', by ascribing such high priority to reliability that the validity of even our claims to having assessed student learning is moribund. Subject profiles provide a means of valuing the full range of assessment practices available to teachers by enhancing their professional responsibilities for valid assessments, within a quality assurance framework, and without sacrificing reliability. (Rowe & Hill, 1996, pp. 339 –340)

The analyses in this chapter suggest there is value in the closer examination of teacher and tests assessments on common scales at the individual student level. Early in the chapter approaches to establish the degree of calibration of teachers to the test scale were described. Scatter plot and 45 degree line comparisons were proposed for individual teachers to establish the degree of calibration and as part of teacher moderation and scale training. Examination of comparisons at the individual student and individual teacher level should lead to greater validity in the assessments from both teacher and test sources. Learning progressions linked to the scales could then help teachers determine the best ‘what next?’ for each student and enhance the professional role of teachers. The more frequently available data points, through teacher judgement, should lead to a greater appreciation of the trajectory of each student.

The next chapter builds on these findings. It considers the time dimension view of learning data and the role of teacher judgement assessments in providing many more data points than are currently possible. Models of learning growth are developed for use in subsequent analyses.

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

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

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

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

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

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

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

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

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

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

Establishing trajectories of learning growth with age and Year level

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

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

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

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

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

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

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

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

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

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

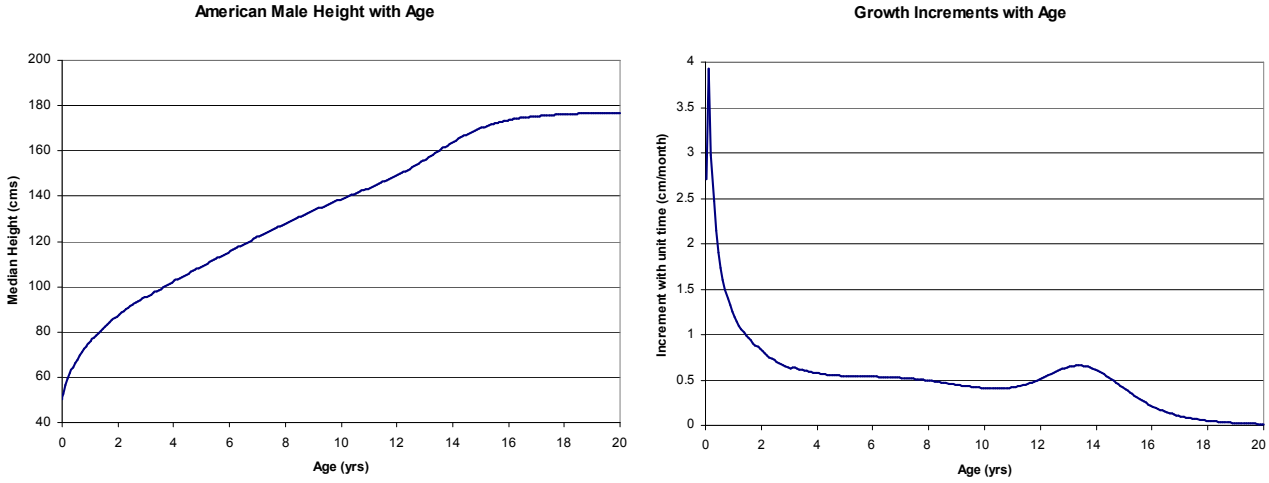
Scale shrinkage, an observed reduction in variance from early in the school year to later in the year and as the Year level increases in some tests (Yen, 1985, 1986), is also seen as a risk to vertical scales (Camilli, 1999; Camilli, Yamamoto & Wang, 1993). For some other tests e.g. the Iowa Tests of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

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

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

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

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



The units for the measurement of learning

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

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

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

Research evidence for unidimensionality across Year levels

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

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

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

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

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

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

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

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

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

Growth in learning status -the time (horizontal) dimension

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

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

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

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

---

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

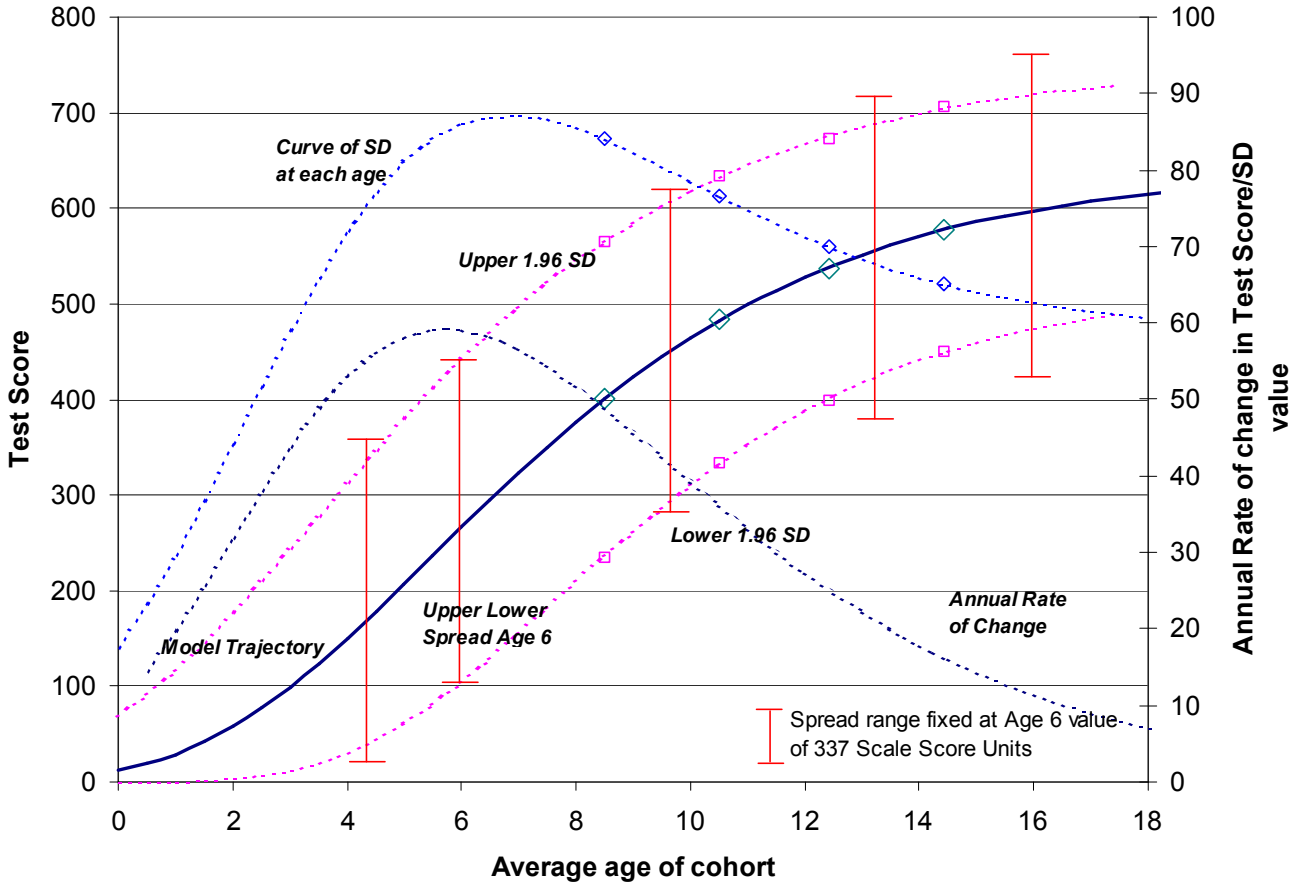
[22] Appendix 5 establishes, as part of the general development of processes to model the general trajectory of learning with time, that comparing Australian system means to an age fitted curve may provide a fairer comparison of performance than the Year level means. System performance is masked when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

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

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

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

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



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

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

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

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

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

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

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

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

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

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

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

Effect sizes for annual growth

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

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

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

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

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

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

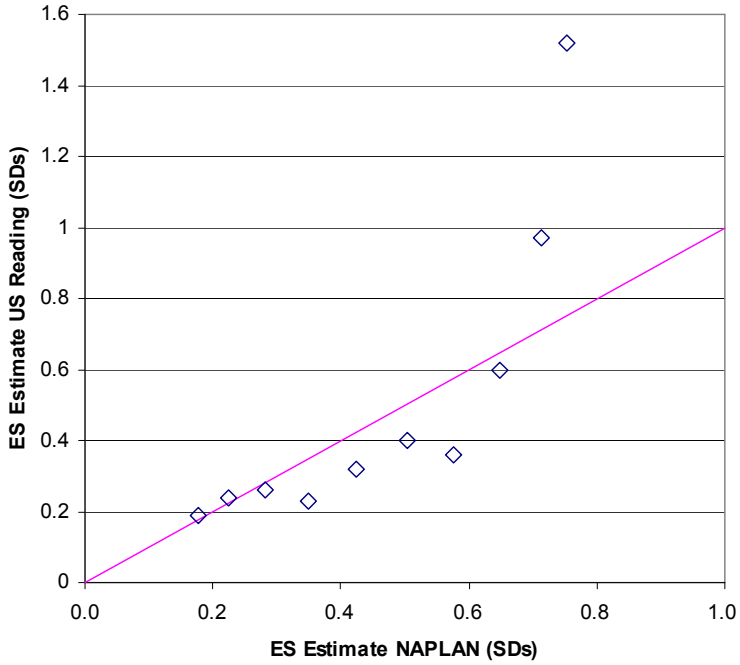
US Data: Table 1, Hill et al., 2007.

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

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

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

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



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

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

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

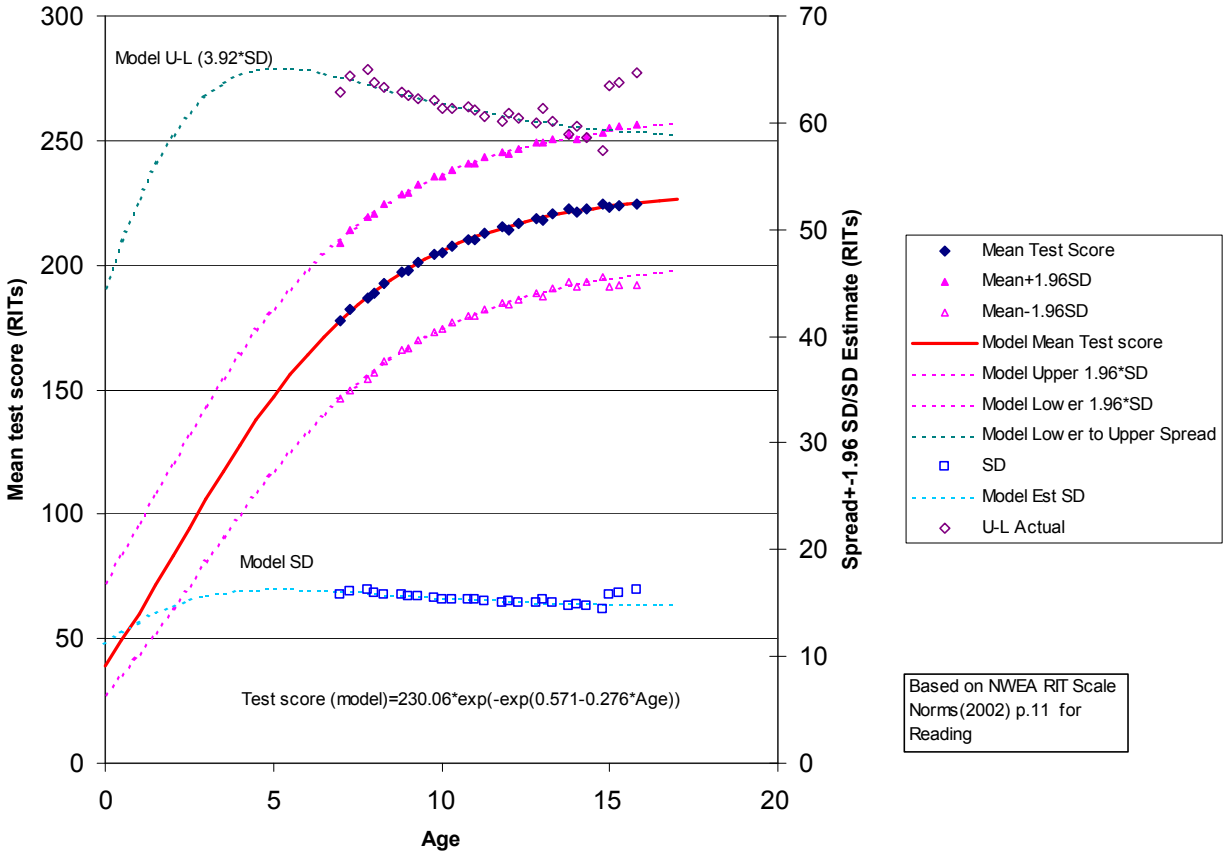
testing program. It is remarkable how a simple and useful model can be estimated from 4 main data points and their SDs. Access to student level records with the ability to monitor the longitudinal growth of individual students from Year 3 to Year 9 should lead to even more refined and interesting models.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

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

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

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

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



Source: Northwest Evaluation Association (2002)

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

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

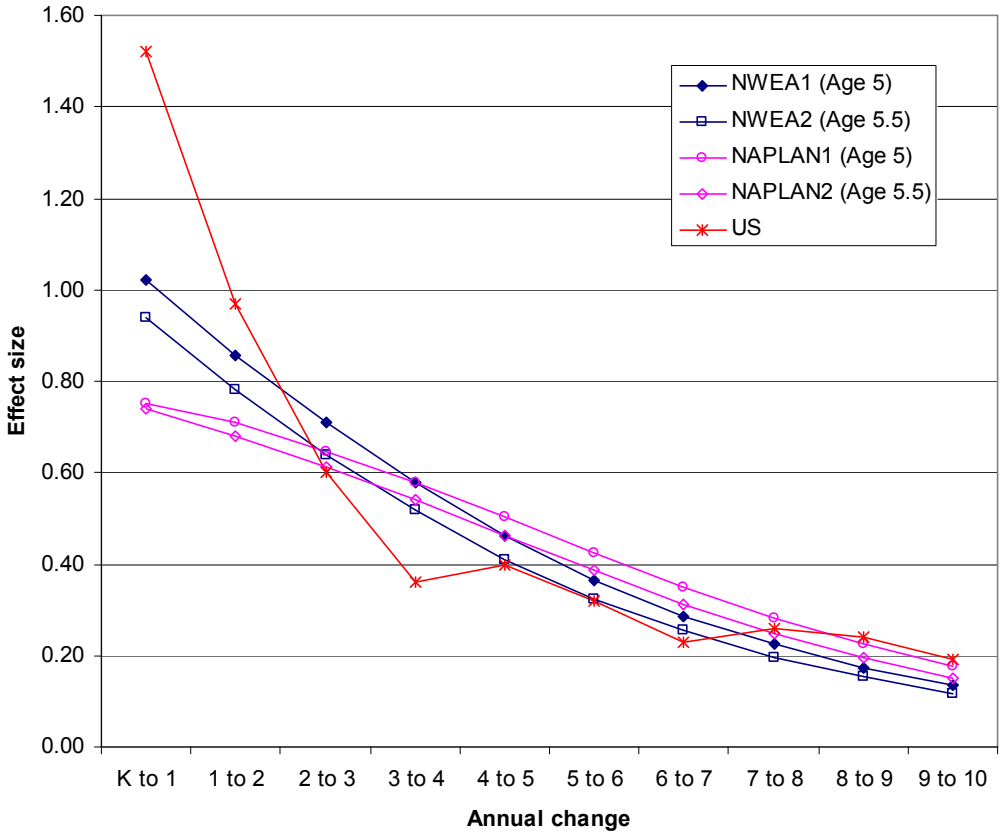
Points for fitting the upper and lower boundary lines are calculated from 1.96 x SD and the curves fitted independently. A model for the 2.5th to 97.5th percentile spread can be developed from the modelled upper and lower boundaries by subtracting the difference for each age point. The resultant Upper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

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

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

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

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



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

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

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

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

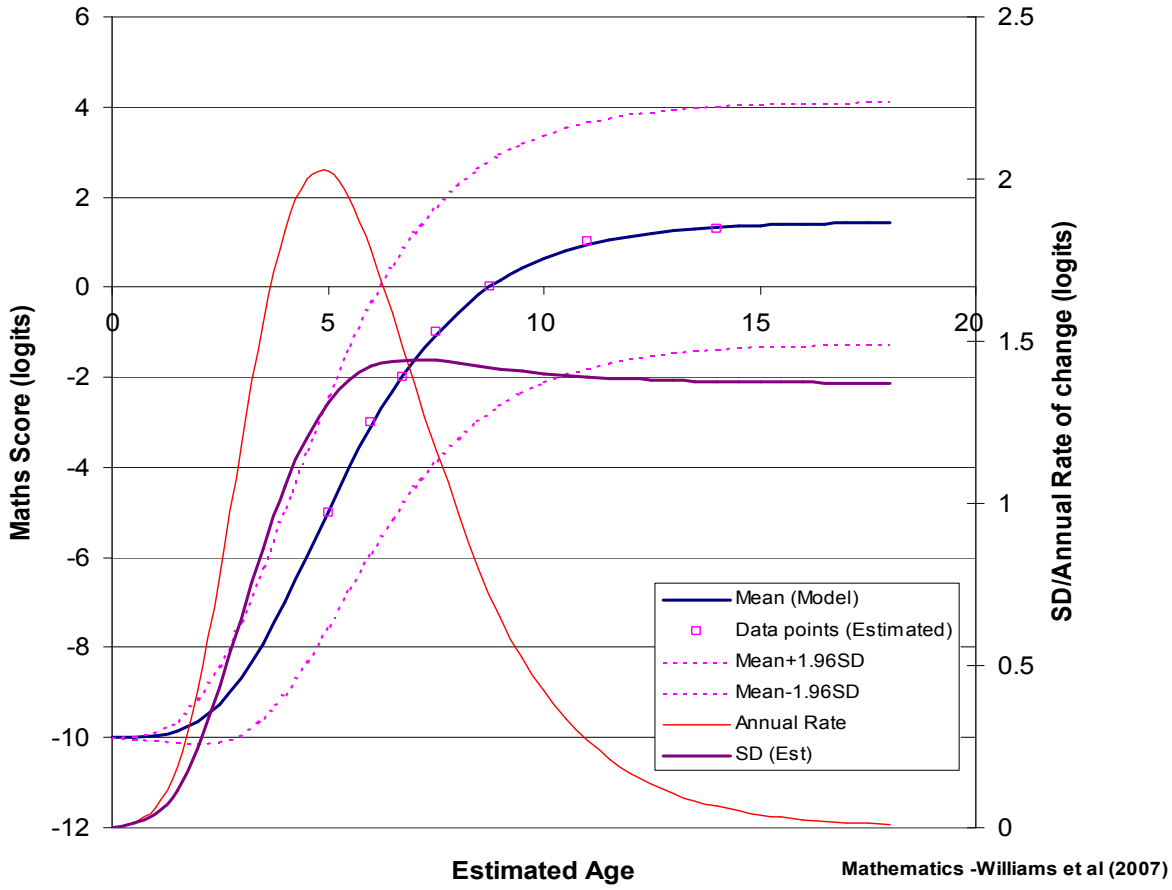
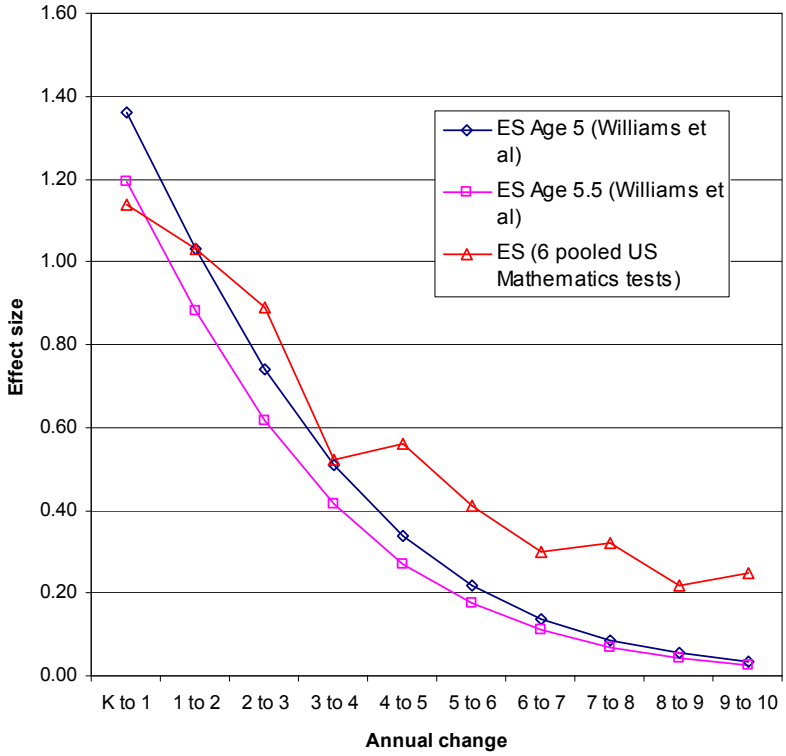


Figure 5.6 displays the resulting model for the mean, the actual data points and the estimated upper and lower boundaries for 90% of the data. Also plotted is the annual rate of change based on the model with its scale on the right hand axis. The estimate of the SD is also

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

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



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

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

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

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

General conclusions on learning growth trajectories from cross-sectional data

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

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

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

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

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

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

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

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

Patterns by age within Year level

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

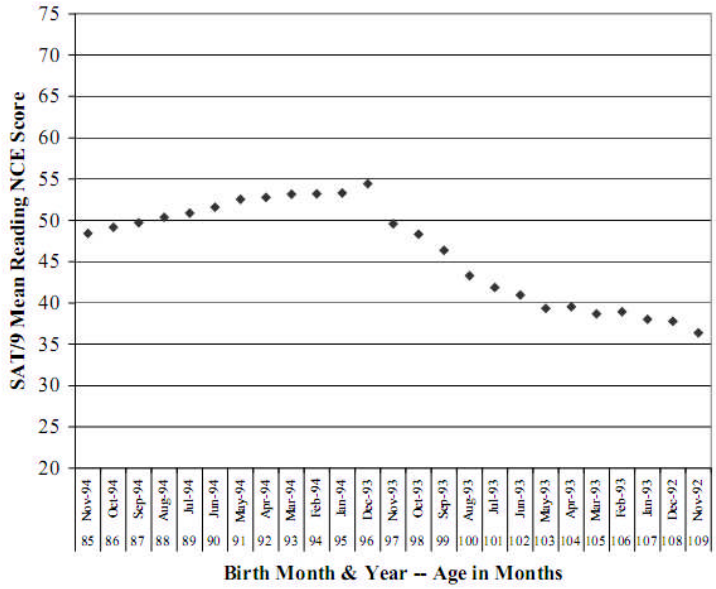
Cahan and Davis (1987) show an increasing percent correct score for reading and mathematics within a cohort with age (in months) for Israeli children. The same test was applied in Year levels 1 and 2, and the increasing score within grade by age was consistent in both grades. The scores for students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

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

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

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

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

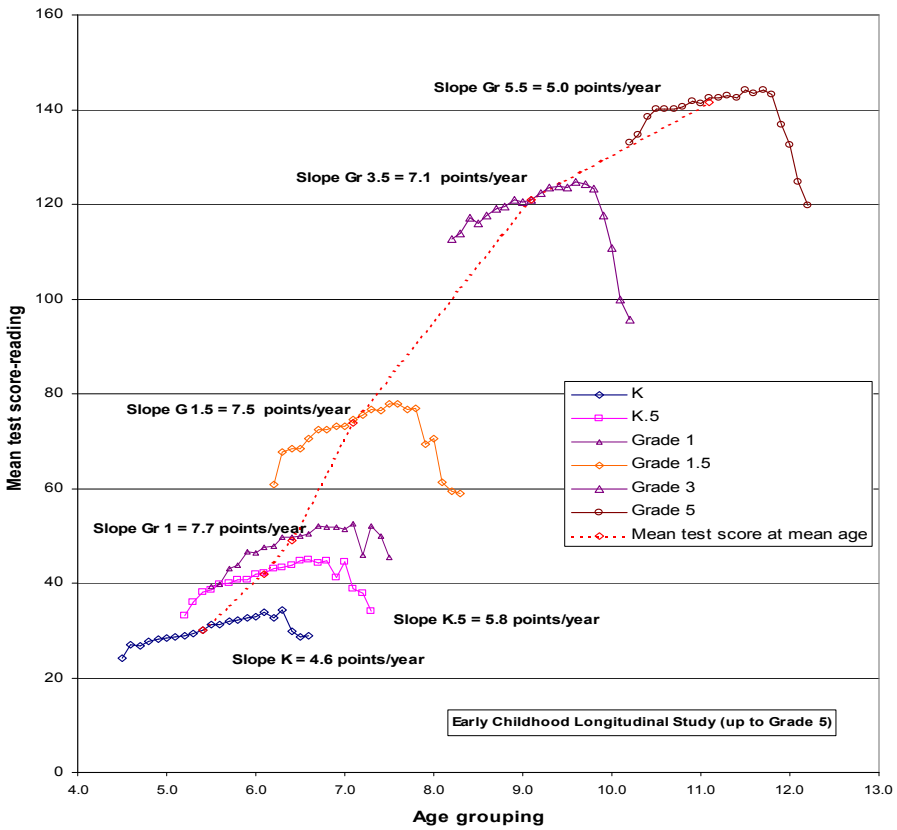


Williams et al. (2007) report the same general learning growth with age within a Year level cohort in the norms developed for the MaLt project described earlier. The gradients within

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

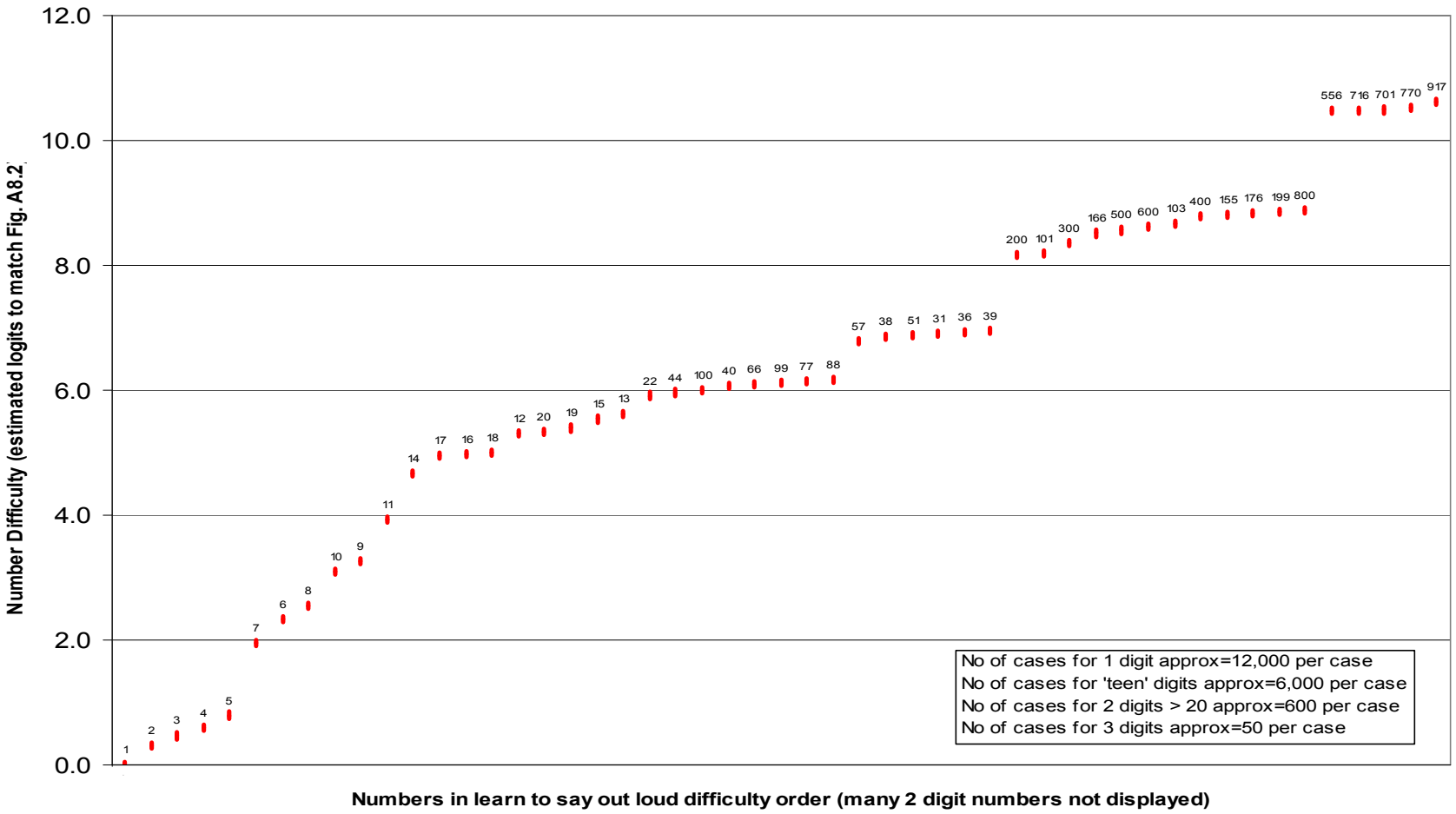
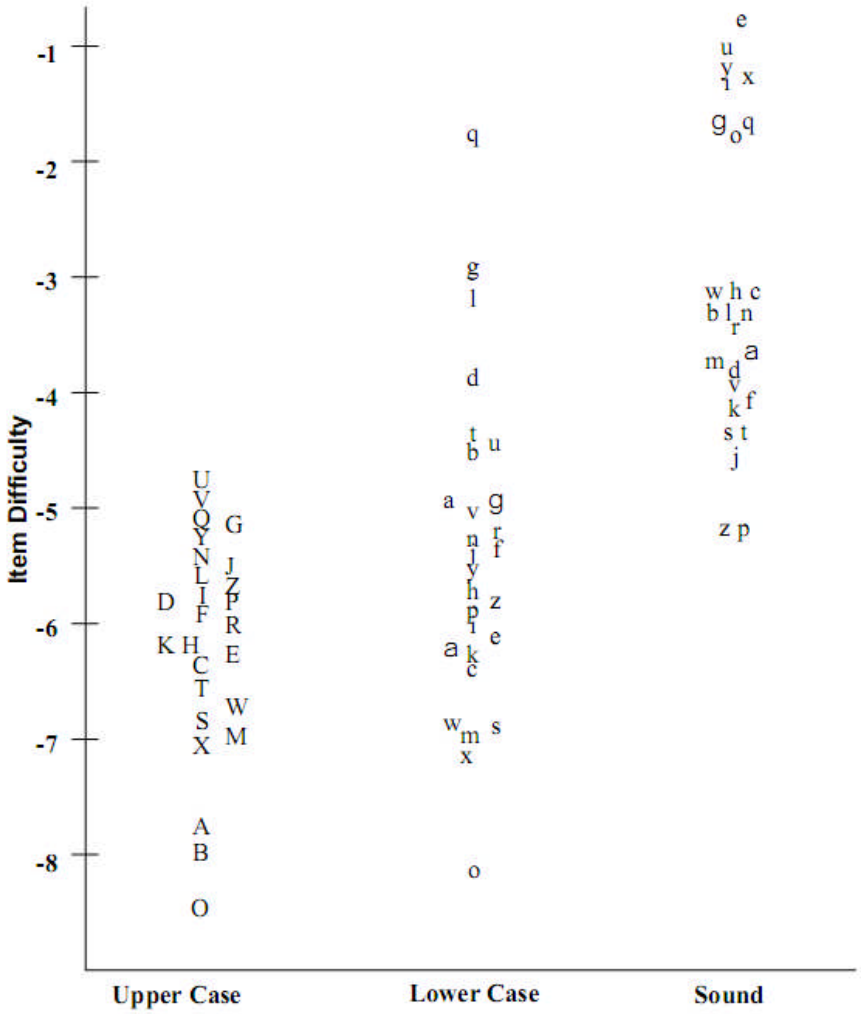


Figure 5.11 Overview of STEP Letter Identification and Letter Sound Item Maps (from Figure 5 Kerbow & Bryk, 2005)



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

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

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

Comments on individual learning trajectories

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

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

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

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

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

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

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

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

Summary

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

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

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

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

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

A wide range of matters relating to testing, teacher judgement assessment, the development of levelled curricula, the application of teacher judgment in schools systems and in this chapter, the patterns of growth that assessment data illustrate have been assembled. These matters set

the context for conclusions that can be drawn in the final chapter of the thesis. The next three chapters analyse and summarise the specific learning trajectories that can be developed from SA test and teacher judgement data of 1997 and 1998. These are developed independently for tests and then teachers and then the two data summaries compared.

Chapter 6: South Australian test data for 1997 and 1998

> A measure of growth is based on measures of status on three or more occasions, obtained either by averaging two or more 'gains' or by modelling growth (curve fitting).
>
> Masters, Rowley, Ainley & Khoo, 2008, p. 16.

The previous chapter explored trajectories of learning. One purpose was to develop a model of test assessment means over a wide range of Year levels to provide a basis for comparison with teacher judgment assessments of students. Teacher judgement assessments, described in the next chapter, provide a consecutive Year level view from Year 1 to Year 8. Test assessment data on the other hand only exist for Years 3 and 5 in South Australia in 1997 and 1998.

To provide a comparable test assessment view over Years 1 to 8, a model of consecutive Year level test assessments is developed in this chapter, extrapolated from known data for Years 3, 5, 7 and 9. These data are drawn from tests for the same student populations, as close to the period of teacher judgement assessment as possible. Data for non-tested Year levels are imputed at a student level using the characteristics of learning growth with Year level and age illustrated in Chapter 5. The extent to which the resulting models represent the real situation is contestable, as for all models. The purpose of the models in this broad analysis is to provide a basis for approximate comparisons of the results of the alternative assessment processes. Accepting the limitations of the models, do teacher and test assessment approaches represent learning development in sufficiently equivalent ways?

The Basic Skills testing program (BSTP) commenced in South Australia with trials in 1994 and full cohort testing for Years 3 and 5 implemented in 1995. A brief history of the implementation of the BSTP is covered in Chapter 3 and is available in more detail in Hungi (2003). These tests have already been part of an extensive publicly reported analysis (Hungi, 2003) covering the years 1995 to 2000. The mean Year level scores for Years 3, 5 and 7 from the tests along with the individual student scores enable a speculative model of individual student scores from Year 1 to Year 8 to be developed. This model, while an imperfect substitute for actual data, provides a basis for a comparison with teacher judgement assessments for the same Year levels. The two data sets are compared later in Chapter 8.

Literacy and Numeracy Tests

The Basic Skills test had two main parts; a Literacy section and a Numeracy section, with subscales within each section. Test items were vertically scaled through common items in the

two Year levels. Student responses were analysed using the Rasch model. Scores were transformed from the original logit scores and reported in a range from 0 to 99, to one decimal point. This thesis uses the original logit scores. As reported in Chapter 3, the South Australian tests were the same tests as used in New South Wales schools for the same years. The NSW Department of Education developed the tests and provided South Australia's results.

The Department of Education and Children's Services (DECS), known as the South Australian Department of Education and Training in 1997 and 1998, approved access to data in February 2005 (Appendix 1). In all DECS made available: individual test records for Years 3 and 5 for the 1997 and 1998, Year 7 records for 2001 (the first year for Year 7 testing), 2002 and 2004. It also made available teacher judgement assessments for 1997 and 1998, reported in Chapter 7. Table 6.1 indicates the number of students included in the main test data files used.

Table 6.1 Students in the Basic Skills Test Program (BSTP) included in data analysis

	1997	1998	2001	2002
Year 3	12437	12794		
Year 5	11973	12471		
Year 7			12873	12930

Test measures of the students in Years 3 and 5 in Literacy in 1997 and Mathematics in 1998 are presented for general comparisons with teacher judgement assessments for these same cohorts. Detail of individual items and the performance of specific items is not a focus. Students' scores are the main interest.

Rasch model analysis of the Literacy and Numeracy tests

To reconcile files provided, to be assured that their structures were fully understood and to re-estimate the error of measurement for each student, a Rasch model analysis using Winsteps was applied to each of the data sets. This was a repeat of the original analysis by the NSW Department. The original data files provided to the author included the Literacy and Numeracy scores for each student on a common scale for Years 3 and 5 in logit form as well as the item responses for each student. They did not include SEs or fit statistics. Details identifying common items in the Year 3 and Year 5 tests were also missing. As a result the Rasch analysis was run for each Year level independently[24] and then crosschecked for consistent results with the originally supplied student logit scores.

[24] Logit scores on the common scale were already known and the only missing details were error of measurement estimates and fit statistics. A request for further information from the SA was deemed

Fit and measurement statistics obtained are tabulated in Tables 6.2 and 6.3. These are presented to confirm the general adequacy of the test. Results from this reanalysis were compared with the original summary scores and the recreated errors of measurement, infit and outfit values added to the record for each student. This was done to establish the individual errors of measurement to be used later in comparisons with teacher judgement assessments.

Table 6.2 Summary of Winsteps Fit and Measurement Statistics, Literacy 1997

Items	N	Measure (mean)	Model 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	1.00	0.00	1.00	40.80	0.02	1.00	0.99	0.12	0.98	0.24
Test Y5	83	0.00	0.03	1.25	0.01	1.00	45.11	0.03	1.25	0.99	0.11	0.96	0.23
Students													
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

Students	Above 1.3 Infit MS	Below 0.7 Infit MS
Test Y3	1.8%	0.1%
Test Y5	2.9%	0.7%

Table 6.3 Summary of Winsteps Fit and Measurement Statistics, Numeracy 1998

Items	N	Measure (mean)	Model 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	32	0.00	0.02	1.19	0.00	1.00	48.52	0.02	1.19	0.99	0.10	1.00	0.18
Test Y5	48	0.00	0.02	1.25	0.01	1.00	48.57	0.03	1.25	0.99	0.07	0.99	0.15
Students													
Test Y3	12794	0.91	0.49	1.25	0.11	0.84	2.27	0.50	1.14	1.00	0.18	1.00	0.50
Test Y5	12471	1.03	0.39	1.10	0.09	0.87	2.55	0.40	1.03	1.00	0.15	0.99	0.44

Students	Above 1.3 Infit MS	Below 0.7 Infit MS
Test Y3	5.6%	2.0%
Test Y5	3.3%	0.8%

unnecessary since the individual student errors of measurement from the individual Year level analyses were assumed to be similar to those found in the common analysis.

Comment on Tables 6.2 and 6.3

The mean of the Year 3 Literacy student scores was 1.03 logits. Effectively the mean difficulty of the test items (0.0) was 1.03 logits easier than the average learning status of the students taking the test, meaning that the test was well targeted, certainly not too hard for the majority of the students. While the original linked analysis placed Year 3 and Year 5 items on a common scale, the relative item placements and spacings of the Year 3 items would be expected to vary only slightly between the linked analysis and the unlinked analysis.

The Year 5 Literacy test was also easier than the average learning status of the students taking the test by 1.42 logits, making it relatively easier for the Year 5 students than the Year 3 test was for the Year 3 students. The test was not too hard for the majority of students. The recreated Year 3 and Year 5 scores from the unlinked analysis scores for each student differed systematically from the original logit scores provided from the original linked analysis. Author-derived Year 3 1997 literacy scores for each student were consistently about 0.67 logits above the original combined Year 3-Year 5 scores provided by the NSW Department, developed using an across-Year level vertical scale.

Year 5 student scores were consistently 0.13 logits below the linked scores. The consistency of relationship at an individual student level between the combined and separate analyses indicates both analyses obtained equivalent scores for students, and that the re-estimated errors of measurement for each student could be assumed to be equivalent to those obtained in the original NSW linked analysis, allowing them to be used in a confidence interval comparison at a later stage in the analysis.

Standard deviations for the distributions of scores were slightly lower for the unlinked analysis than for the linked analysis (1.30 compared with 1.36 for Year 3, 1.22 compared with 1.24 for Year 5). This difference is due to the wider range of student scores in the combined analysis. That the difference is so small indicates that the range of scores for each Year level separately is almost identical to the combined range. Similar patterns applied for Numeracy in 1998 (Table 6.3), with the standard deviations increasing slightly in the linked analysis.

For subsequent summaries and analyses the original student scores from the linked scales analysis were used, with re-estimated errors of measurement and fit statistics added to each student record.

The mean score differences (that is growth) for the original logit scores were 1.19 logits (Year 3 mean 0.36, Year 5 mean 1.55) for Literacy in 1997, and 1.21 logits (Year 3 mean 0.13, Year 5 mean 1.34) for Numeracy in 1998. Hungi (2003) applied an analysis with a more sophisticated equating and linking process than generally applied by NSW analysts. He established that a linked analysis over the calendar years (1995 to 2000) varied the original

mean scores at each Year level and for each calendar year by up to 0.05 of a logit and the resultant growth estimates from Year 3 to Year 5 by up to 0.1 of a logit (see Table 6.4 and 6.8). This variation is equivalent to about 1/5th of a year's learning and adds an additional tolerance consideration when test and teachers judgement assessments are compared.

The fit of Literacy items in 1997 and Numeracy items in 1998 to the Rasch model was good as shown in Tables 6.2 and 6.3. Infit mean square values for the items were close to 1.0 with none outside the range 0.7 to 1.3. The tables indicate the percentages of students with infit values above and below the values of 0.7 and 1.3. Cases below 0.7 are negligible except for Year 3 Numeracy where 2% are estimated to overfit implying a small degree of item dependency for these students. Misfitting students (above 1.3 Infit values) are between 2% and 3% of cases, except again for Numeracy in Year 3, where almost 6% of students misfit.

Item reliabilities are consistently 1.0 (due to the very large N of students tested). Reliability from a student measurement perspective is less for the Numeracy test (0.84, 0.87) than for the Literacy test (0.91, 0.92). The mean model error of measurement for students is of the order of 0.3-0.4 logits (0.49 for Year 3 Numeracy) implying an error in estimating individual student learning status of up to 8 months learning development. This error is greater for Year 3, most likely reflecting the complexities in measuring numeracy skills for students with low numeracy and most often low reading skills with a self completed paper and pencil test. The high percentages of students with infit values above 1.3 infit mean square in Year 3 (Table 6.3) supports this explanation.

At this stage in the data development, of the order of 12,000 records for Year levels 3 and 5 (as shown in Table 6.1) are available with student scores, errors of measurement and fit statistics. These records serve two purposes. About 1000 cases per year level have a potential match with the teacher judgement assessments to be taken up in Chapter 8. The second purpose is to contribute student cases to the development of model data sets using the understanding of trajectories of learning from the previous chapter to impute values for notional students in Year levels 1 to 8. This is done to provide a comparison with the teacher judgement assessments. If students had been assessed by tests at all year levels what might that data have looked like?

The trajectory of Literacy test scores

To aid in the estimation of the trajectories of learning as Year level increases, a wide range of South Australian data are reviewed. These data are considered as part of the process to select data points for the mean scores in Literacy at each of the tested Year levels. A curve is fitted to these points using the Gompertz expression as described in Chapter 5. This trajectory then becomes a framework for estimating the means for missing Year levels. Data for typical

students at each missing Year level are then imputed. This is done by adjusting the scores for a random sample of students drawn from the tested Year levels (3, 5 and 7) so that the mean scores approximately match the framework. Adjacent cohorts (3 and 5, 5 and 7) contribute equally to the samples for Years 4 and 6. Cohorts below Year 3 are 'stretched' to ensure that both the overall means and the means by 0.1 of age follow the framework trajectory. This is achieved by including the age at testing as well as the Year level for each imputed student. The age at testing values for each student then allow a more general set of summaries of the model data to be made.

Table 6.4 summarises the mean scores in Literacy for SA students by tested year level from 1997 to 2004 as well as NAPLAN data for 2008. Data for 2003 were not in a form that could be readily summarised and are omitted. Four perspectives of the data are provided, two cross-sectional and two longitudinal. This is done to establish that all four perspectives generate essentially the same curve of learning development with Year level and age, confirming the Hilton and Patrick (1970) finding that the general change from testing period to testing period was similar whether the group of interest was cross sectional or longitudinal-not matched (the cohort not adjusted for losses and gains).

The first cross-sectional block of Table 6.4 (1997 as an example) presents the original SA cross-sectional values at Year 3 and Year 5 (and Year 7 in some cases). The range of scores within a Year level is wide (0.18-0.61 at Year 3, 1.38-1.79 at Year 5). Growth values are in a narrower range (1.03 to 1.23 logits for Year 3 to 5, 0.70 to 0.89 for Year 5 to 7).

The second cross-sectional block (1995, Hungi adjusted) is the result of Hungi's re-scaling based on a multi-linked analysis. The mean values at each Year level estimated by Hungi are not strictly comparable to those values originally derived for each calendar year. The value of 0.31 (1997 Year 3 as adjusted by Hungi) on the common item scale for example, may not be positioned at exactly 0.31 on the original scale, since Hungi refined the scale to be better calibrated across the testing years 1995 to 2000 than was originally developed in NSW. An implication of the Hungi analysis is that the NSW item scale had the potential to vary from calendar year to calendar year. The growth estimates at the right side of the table represent differences between scale values and should be more comparable even if not in identical units.

The third block presents a longitudinal view (cohort wave identified by the calendar year at Year level 3) showing the values for the Year 3 cohorts at successive 2 yearly intervals. As an example the Year 3 value in 1997 is related to the Year 5 value in 1999, two years later.

The fourth perspective is also longitudinal, similar to the third, but uses the re-calibrated Hungi values to indicate growth values based on the same cohorts two years apart.

Table 6.4 Literacy – Mean scores by Year level and Testing Year

	Year Level	3	5	7	9	Growth	Growth	Growth
	Average age	8.6	10.6	12.6	14.5	3 to 5	5 to 7	7 to 9
Cross-sectional	1997	**0.36***	**1.55***			1.19		
	1998	0.18	1.42			1.23		
	1999	0.44	1.47			1.03		
	2000	0.30	1.38			1.08		
	2001	0.45	1.60	2.30[25]		1.15	0.70	
	2002	0.39	1.61	2.41		1.22	0.80	
	2004	0.61	1.79	2.68		1.18	0.89	
	Mean	0.39	1.54	**2.46***		1.15	0.80	
	1995 (Hungi adjusted)	0.38	1.36			0.98		
	1996 (Hungi adjusted)	0.30	1.48			1.18		
	1997 (Hungi adjusted)	0.31	1.57			1.26		
	1998 (Hungi adjusted)	0.22	1.50			1.28		
	1999 (Hungi adjusted)	0.38	1.31			0.93		
	2000 (Hungi adjusted)	0.18	1.29			1.11		
	Mean	0.30	1.42			1.12		
Longitudinal	1997 Cohort wave	0.36	1.47	2.30		1.11	0.83	
	1998 Cohort wave	0.18	1.38	2.41		1.19	1.04	
	1999 Cohort wave	0.44	1.60			1.16		
	2000 Cohort wave	0.30	1.79	2.68		1.49	0.89	
	Mean	0.32	1.56	2.46		1.24	0.91	
	1995 Cohort wave (Hungi)	0.38	1.57			1.19		
	1996 Cohort wave (Hungi)	0.30	1.5			1.20		
	1997 Cohort wave (Hungi)	0.31	1.31			1.00		
	1998 Cohort wave (Hungi)	0.22	1.29			1.07		
	Mean	0.30	1.42			1.12		
NAPLAN	NAPLAN SA 2008 Reading (estimated logits)	0.40	1.51	2.30	**2.89***	1.11	0.79	0.59
	NAPLAN SA 2008 Reading (reported scores)	*(400.5)*	*(477.9)*	*(533.5)*	*(575)*	*(77.4)*	*(55.6)*	*(41.4)*

* Bold values signify values used in estimation of Gompertz model parameters.

The final block of the table indicates the reading data for SA's 2008 NAPLAN tests. This is not the same as the more general Literacy applying in 1997 but is assumed to be a reasonable indicator of the general trajectory. These data are tabulated to provide an additional influence on the model trajectory for Years 7, 8 and 9 developed below. The original scores are provided, along with a conversion on an estimated basis to logits. Growth from Year 3 to Year 5 over a number of years is estimated to be approximately 1.12 logits based on the grand mean of Hungi's estimates. The growth of 77.4 NAPLAN units is used to estimate an approximate conversion factor of 70 units to 1 logit. On this basis, after setting the NAPLAN

[25] The estimates for the means at Year 7 are derived from the original files provided. About 300 cases were omitted for 2001, 50 for 2002. These cases had the lowest possible scores for Literacy but high scores for Numeracy. It is assumed that these were cases with no data for the specific test and allocated minimum scores in the original analysis. The author's summary omits them.

value at Year 3 as 0.4 logits (to fix it close to the Year 3 1997 Literacy value), approximate estimates in logits can be made for the reading means at Years 5, 7 and 9. The resultant growth values are comparable to the other growth trends in the table[26].

In summary, the growth in mean literacy learning over a mix of years and cases when viewed as cross-sectional is very similar for the original SA data and for the Hungi re-scaling. For longitudinal growth, the Hungi re-scaling is also similar to the cross-sectional growth. The growths in the longitudinal view of the original data are also similar, although the mean is possibly skewed by the 2000 Cohort wave value. This is illustrated graphically later in the chapter and is possibly an artefact of less accurate across-calendar year equating that Hungi has addressed in his re-equating of the scales. Overall this extensive analysis of the change of scores from one tested Year level to the next tested Year level establishes that the patterns of group mean growth in scores are essentially the same whether a cross-sectional or longitudinal view is used. The period of interest (1997) very conveniently approximates the mean values for all the cases examined. The 1997 values plus the NAPLAN extension to Year 9 are used in the next section to develop a framework for a model of literacy growth in South Australia in 1997.

Developing a model for the test trajectory of learning – Literacy.

The model is developed in two stages. Initially a general model for the mean learning status at each year level is developed: the framework. In the second stage, data points for individual students at each missing year level are created, based on actual student data from Years 3, 5 and 7. A number of assumptions are made in the model development.

Patterns of mean learning development by Year level are assumed to be similar over different calendar years. While these patterns vary, evidence from Chapter 5 (and Tables 6.4 and 6.8) indicates that these patterns are consistent enough to provide a trajectory framework for group means. It is assumed the location of the mean for an intermediate Year level (Year 4, Year 6) can be placed on a smooth curve describing the trajectory, accepting the Year level units as equally spaced time units on the X-axis. In the absence of actual SA test data describing the trajectory of literacy learning leading up to Year 3, it is assumed that the means for Year 1 and 2 can be placed on a smoothed trajectory, using a Gompertz expression as outlined in

[26] The average age for each cohort has increased in the period from 1997 to 2008. Based on estimates from the annual age and Year level census bulletins (ABS 4221), the average age at July 1 has increased by about 0.15 of a year of age at Year 1 and by 0.1 at Year 7. The testing period has also shifted to earlier in the school year (August in 1997 to May in 2008), meaning that the age at testing has not varied much over this period.

Chapter 5 to estimate the trajectory. As illustrated in that chapter, when the NAPLAN model is compared to a number of US normed tests, a steeper growth curve in the early years of school is found for the US data. For Literacy the Gompertz model trajectory for SA data therefore may not be steep enough. The means for Years 1 and 2, derived from the Gompertz curve fitted to the data in Table 6.4, may be conservative in the degree of learning growth they indicate.

Even if conservative, a steeper trajectory for learning applies in the early years relative to later years and the within-Year level by age curve has a steeper gradient in the lower years. Noting this age effect described in Chapter 5, the spread of scores by age within Years 1 and 2 needs to be increased to reflect this effect. An extensive record matching process to add dates of birth to the student score files is required. The age for each student at testing can then be calculated. While Year 7 scores for 1997 do not exist, it is assumed that the records of approximately 26,000 students (see Table 6.1) in 2001 and 2002 are indicative of the general spread of student scores at this higher Year level. These records are used to add Year 7 and Year 8 cases for the model.

Finally it is assumed that the Year 7 mean scores, combined with the general estimates from the SA NAPLAN data for Year 9, can be used as points in the model to influence the trajectory fitted for higher Year levels. Assuming that the growth from each tested Year level to the next is consistent over testing periods is only partly valid. The larger the test population, the more likely the growth in learning remains consistent. England and US national data reviewed in Chapter 5 show very small variation in mean scores over time. As the unit of analysis becomes smaller, at a school or classroom level, much greater variations in the scores are observed. These variations appear to balance each other out in the aggregated summaries.

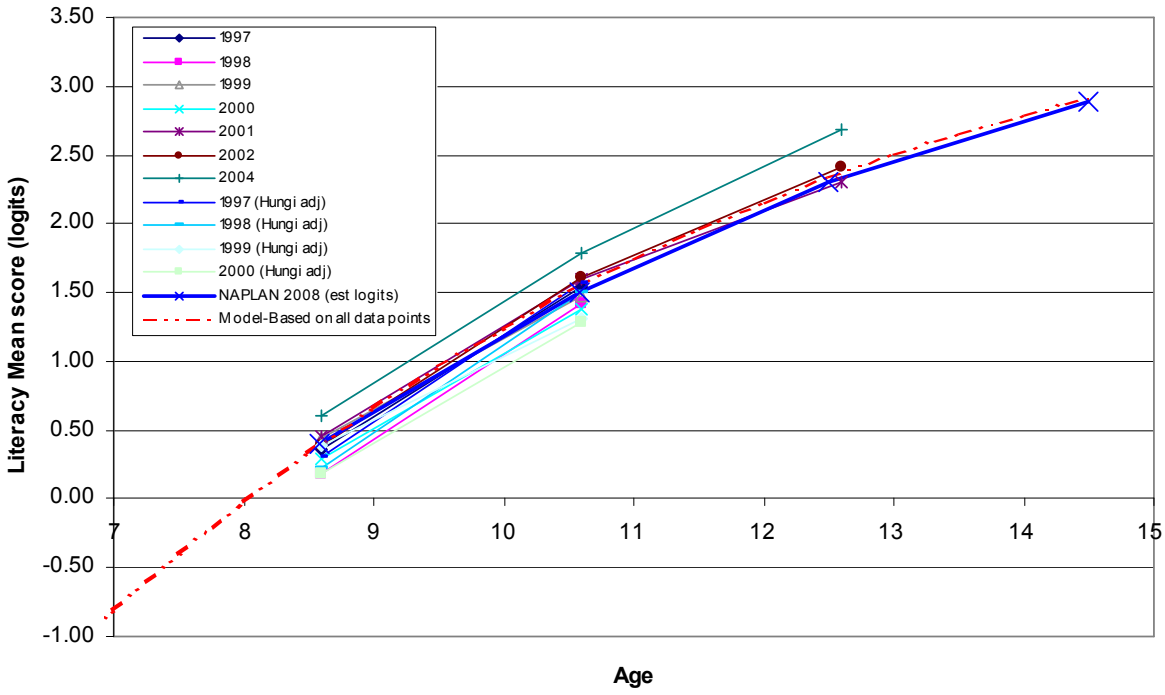
It is recognised that these assumptions are contestable. However a theory on what comprehensive test data might have looked like provides a comparison with the actual teacher data presented in detail in Chapters 7 and 8. Applying the above assumptions about how missing test data should look enables the development of the imputed data in the following sections. First frameworks for the group means for literacy and numeracy are established. Then typical cases are added for each missing Year level using combinations of actual students from the tested Year levels.

Setting the framework for the Literacy model

Figures 6.1 and 6.2 indicate two views of the data in Table 6.4. In both figures the X-axis is converted to age, with the Year level points plotted at the average age for the Year level. In Figure 6.1 the cross-sectional points within a calendar year are connected. On the whole their

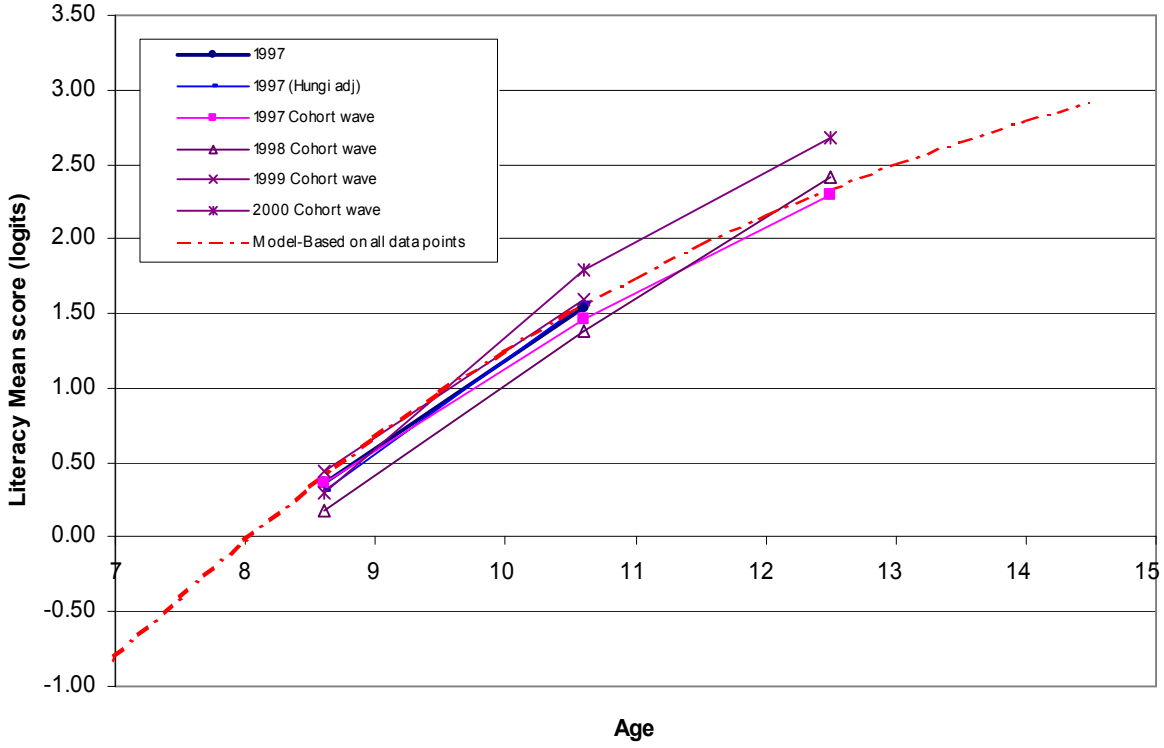
trajectories are similar even though displaced due to variability in the scaling and/or actual variation in the performance of the groups. The line for 2004 shows the same cross-sectional growth pattern as the rest but is displaced above the group. This is possibly related to a change in the testing and analysis provider and the resultant difficulty in aligning with the previous scale. The SA NAPLAN reading data (as distinct from Literacy) appear to follow the same general trajectory.

Figure 6.1 Literacy mean scores –Cross-sectional view with model trajectory



In Figure 6.2 the plotted lines link cohorts at two-year intervals. These follow less consistent trajectories. The greater variability in trajectory for these waves suggests that some variability is due to the variations in the scale with calendar year. Within calendar year linking appears more reliable than the across calendar year linking. Longitudinal views based on Hungi's re-scaled data follow the same general trajectories as cross-sectional data, suggesting that the variation in the longitudinal view of original scores is due less to variation in growth than variation in the scaling process.

Figure 6.2 Literacy mean scores –Longitudinal view with model trajectory



The dashed line is the trajectory obtained using the four points, including SA NAPLAN data, and fitting a Gompertz curve using CurveExpert (Hyams, 2001). To achieve this the raw logit values are scaled upwards by 10 logits to remove all negative values and to place the likely mean learning score value at age=0, very low. This is done, as the Gompertz expression cannot be fitted to negative values. Once a curve is fitted the values for each age point are then rescaled back to the original logit scale origin by subtracting 10 units[27]. Target means can then be identified for each Year level, using the fitted Gompertz curve. The target is calculated using the average age for the Year level. The derived curves are shown in Figures 6.1 and 6.2.

The next stage in the model development is to generate student records for each Year level using sampling of data points from the known data distributions for Years 3, 5 and 7.

Adding multiple points to Literacy model

Imputed data points for each Year level were developed in stages. The first stage required establishing the date of birth for as many records as possible. The original testing process did not collect date of birth, only conventional integer age at the point of testing. The lack of age

[27] The values of the parameters for the fitted Gompertz curve are a = 14.05, b = 0.65, c = 0.22. The estimated score value at any age value is converted to the original logit scale by subtracting 10. The asymptote (a) of the group mean is effectively 4.05 logits on the original scale as age moves above 20.

detail made an analysis of test data by actual age at testing impossible. To remedy this a date of birth was found for as many students as possible. This was a long process of matching names to a master file of names and obtaining the date of birth from the student file along with a unique student identification number. About two thirds of each Year level (8000 of 12000 records for each of four cohorts) were assigned a date of birth. The student identifier was required for later matching to teacher judgement assessments in Chapter 8.

The statistical characteristics of the original files and the subset that were assigned dates of birth are tabulated in Table 6.5. For Year 3 the mean of the sample with birth dates (8988 out of 12437: 72%) has a mean of 0.46 logits, 0.1 logit greater than the full cohort mean of 0.36. This indicates a slight bias in name matching against finding some lower scoring students. The standard deviation, inter quartile range, skewness and kurtosis values of the sample and the original cohort are similar enough to assume that they have similar distributions even though the mean is greater. In the model building process, the sample is set to a new mean by adjusting the value for each individual case by the amount required to make the model mean match the target mean from the framework.

For Year 5, 8651 out of 11972 records (72%) were matched. The mean for the sample with birth dates was almost identical to the full sample.

Table 6.5 Literacy-Comparison of original records with subsets assigned dates of birth

	Sample with Birth Dates		Full cohort	
	Year 3	**Year 5**	**Year 3**	**Year 5**
Mean	0.46	1.56	0.36	1.55
Median	0.51	1.69	0.41	1.69
SD	1.44	1.20	1.36	1.24
Skewness	-0.30	-0.70	-0.34	-0.79
Kurtosis	4.35	5.43	4.42	5.65
IQR	1.74	1.54	1.91	1.61
N	8988	8651	12437	11972
% with DOB			72.3%	72.3%

Once dates of birth were established, the ages of students were calculated and categorised for specific age categories relative to the middle of August 1997, the test period. Age was represented by actual age at testing, in categories of integer age (age last birthday), age in half years, age in 0.2 of a year and age in 0.1 of a year, approximating an age in years and months. Students were placed into the categories on the basis of the relationship of their actual age to interval boundaries. The 0.1 categories were centred on the required values with boundaries at 0.05 of a year. The 0.2 categories were centred on the odd values (0.9, 0.1, 0.3, 0.5 etc) with the even values being the interval boundaries.

Files for Years 3 and 5 were randomly sampled to select 7950 records for each Year level from the larger samples (8988 and 8651 respectively). This limit was required to ensure that the final model of records would fit into the version of Excel being used, which had an upper limit of 64,000 records, allowing a maximum of 8000 cases for each of Years 1 to 8. The mean of the Year 3 sample was adjusted slightly to match the target set in the framework model.

For the Year 4 component the original Years 3 and 5 cases with assigned dates of birth were sampled to select 3975 records from each source. These records were then summarised to obtain the initial Year 4 sample mean. These new data were then adjusted by the required amounts to set the grand mean to the framework target for Year 4. A common amount was added to the each of the Year 3 derived records and a common amount subtracted from each of the Year 5 derived records.

Year 7 data from 2001 and 2002, shown in Table 6.6, included dates of birth as part of the data collection and testing procedure. Cases from both 2001 and 2002 files were combined.

Table 6.6 Comparison of 2001, 2002 and 2004 Literacy score statistics - full cohorts

Statistics	2001 Y7 Literacy	2002 Yr 7 Literacy	2004 Yr 7 Literacy
Mean	2.30	2.41	2.68
Median	2.30	2.46	2.71
Skewness	-0.04	-0.26	-0.26
Kurtosis	3.40	3.54	3.45
SD	0.92	1.08	1.15
Inter Quartile Range (IQR)	1.17	1.43	1.48
SE (Mean)	0.01	0.01	0.01
N	12533	12069	15628

The grand means of two independent samples of the Year 7 composite records were set to the framework model targets by systematic individual record adjustment to generate records for Years 7 and 8. A problem was discovered later, well after the model data were developed. Students who missed one or other of the tests in 2001 were assigned the minimum possible score rather than deleted. This influenced the means. On discovery of the problem zero score records were deleted from the samples, leading to final samples being less than the intended 7950. Year 6 records were developed in the same manner as Year 4 records. Samples were drawn independently for half the required records from Year 5 and Year 7, and each subset adjusted to average to the overall framework required Year 6 mean.

The creation of records for Years 1 and 2 required one additional step. Both were based on independently sampled Year 3 subsets of 7950 records from the Year 3 records with dates of birth. The data were then summarised by 0.1 of an age and the notional mean scores for each

0.1 age category spread down the age scale to follow, in general terms, the steeper gradient at these age points in the model. The adjustments for Year 1 were greater than the adjustments for Year 2 and are justified on the basis of the age within grade/Year level observations described in Chapter 5. Evidence from Chapter 5 suggests that the within-Year level gradient with age does not match the general across age gradient. The within-Year level gradient gets flatter with increasing age but in the first years of school almost follows the general trajectory. The model records were adjusted to show this steeper gradient within Year level. This was done by setting the mean for each 0.1 age cohort to sit on the general trajectory line while keeping the mean for the full Year level cohort at the framework specified value. Records were adjusted manually and iteratively for each 0.1 age cohort leading to approximate matches only of the means to the target for each 0.1 age cohort. The matches of the Year level means at Years 1 and 2 to the framework overall are very close as shown in Table 6.7.

The effects of the score adjustments for each Year level in the model, relative to the target values to be achieved, are shown in Table 6.7. The general characteristics of the final 63,306 simulated students are illustrated. The framework target means and the means of the imputed points for each Year level match well. The inter-quartile range reduces with increasing Year level, as does the SD, as expected from Chapter 5. The SDs in Years 7 and 8 are effectively the same as the two samples are clones from the same Year 7 distributions.

Table 6.7 Literacy Model-main statistical characteristics

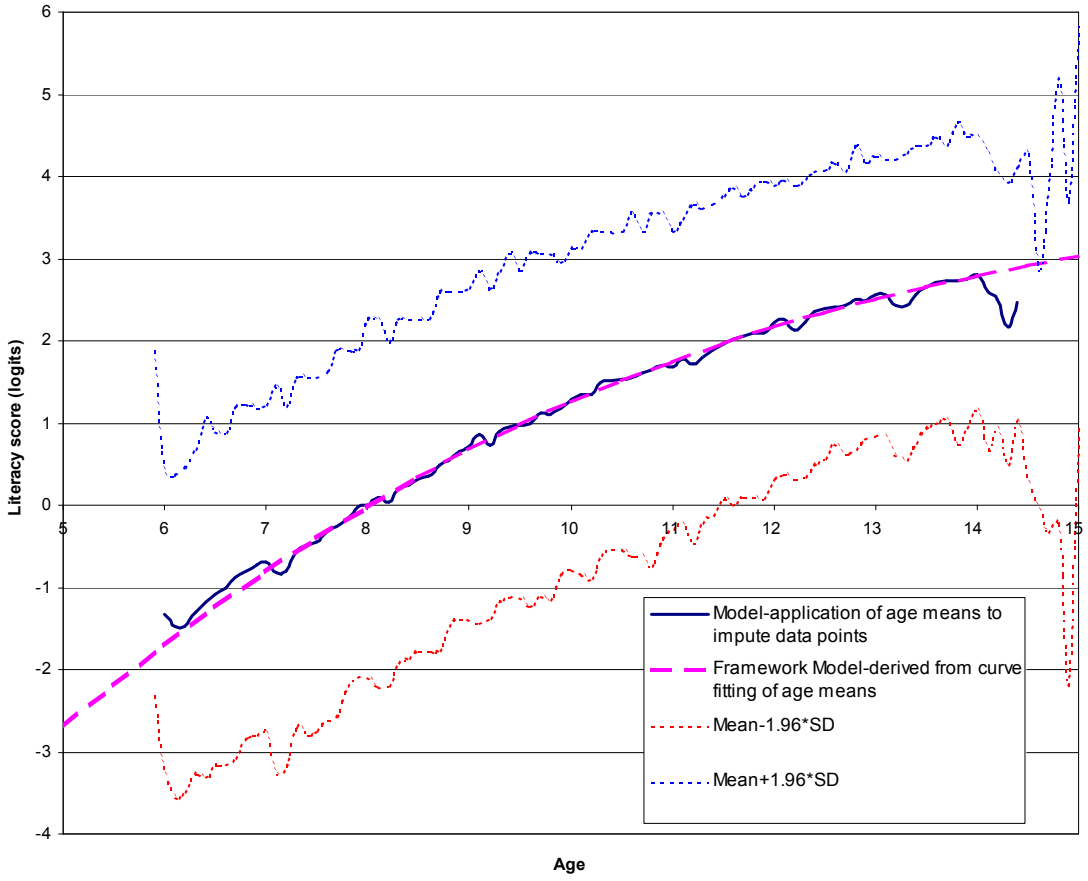
Literacy Model by Year Level	1	2	3	4	5	6	7	8
Framework (Targets for each YL)	-1.14	-0.31	*0.41**	1.04	*1.57*	2.01	2.39	2.69
Means of used or imputed points	-1.14	-0.32	*0.41*	1.04	*1.57*	2.02	2.39	2.70
Medians of used or imputed points	-1.10	-0.25	*0.47*	1.12	*1.69*	2.05	2.40	2.71
SDs of used or imputed points	1.35	1.32	*1.31*	1.26	*1.20*	1.07	1.00	1.02
Skewness of used or imputed points	-0.27	-0.32	*-0.29*	-0.45	*-0.63*	-0.26	-0.15	-0.20
Kurtosis of used or imputed points	4.12	4.42	*4.32*	4.70	*5.03*	3.67	3.63	3.77
IQR of used or imputed points	1.85	1.74	*1.74*	1.65	*1.54*	1.36	1.29	1.29
Count of used or imputed points	7949	7949	*7949*	7949	*7949*	7888	7837	7836

* Italics signify Year levels with actual data, although case values have been adjusted to average to the framework Year level means.

The Model data compared to the Framework

The complete data set for the imputed data points is summarised in Figure 6.3. The data are shown as if they are continuous but are discrete means at 0.1 of an age. The means of the scores at each age point, when calculated independently of Year level, follow a trajectory with age that matches the framework model. This is unsurprising at the lower Year levels since the data were adjusted to achieve this result. However from Year 3 onwards the trajectory of the means is determined by the natural elements of the data.

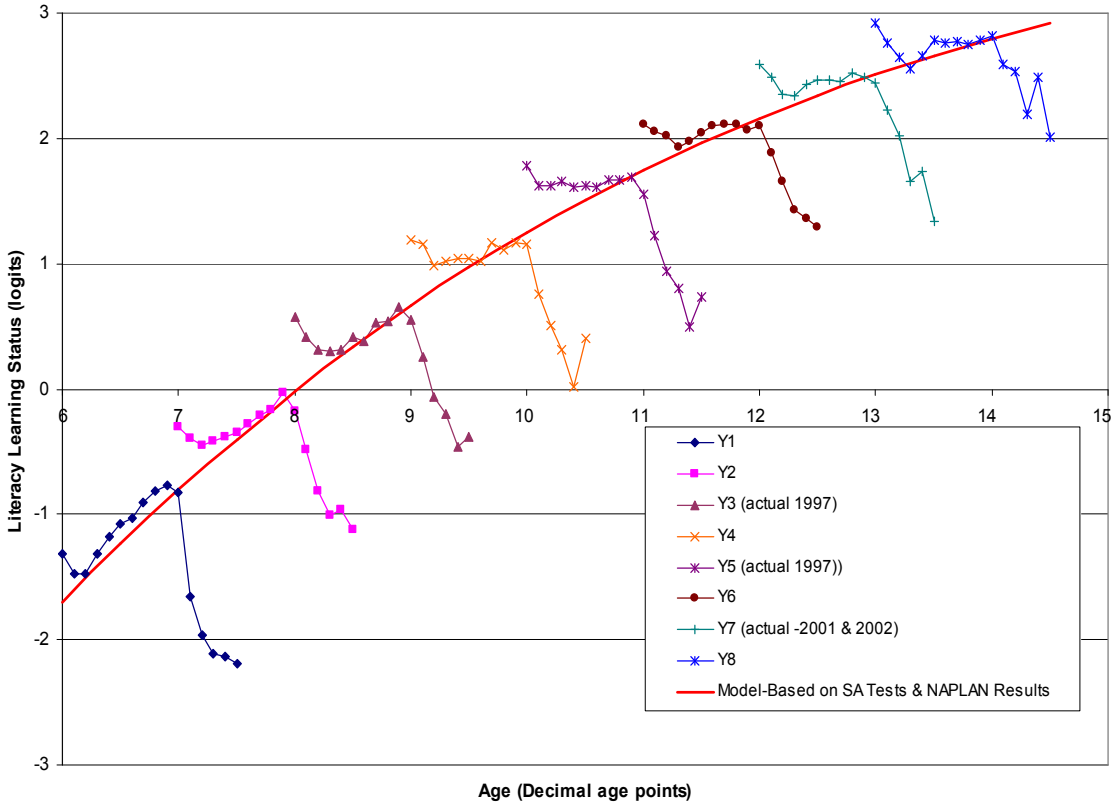
Figure 6.3 Comparison of Literacy Model to the Framework Model



The 2.5th and 97.5th percentile ranges are shown, to indicate that the general pattern with age applies across the spread of the data. Widening the age category (using 0.5 of an age category relative to 0.1) can smooth the fluctuations around the general trajectory but hides the elegance of the pattern with age.

When model data are analysed by age within Year level, the general trend of increased mean score with age within a Year level is revealed, as shown in Figure 6.4. Once again the trajectories at Years 1 and 2 are artificial. From Year 3 onwards these reflect the patterns in actual data. The effects at Years 4 and 6 are achieved by blending years above and below, which may not reflect actual data patterns.

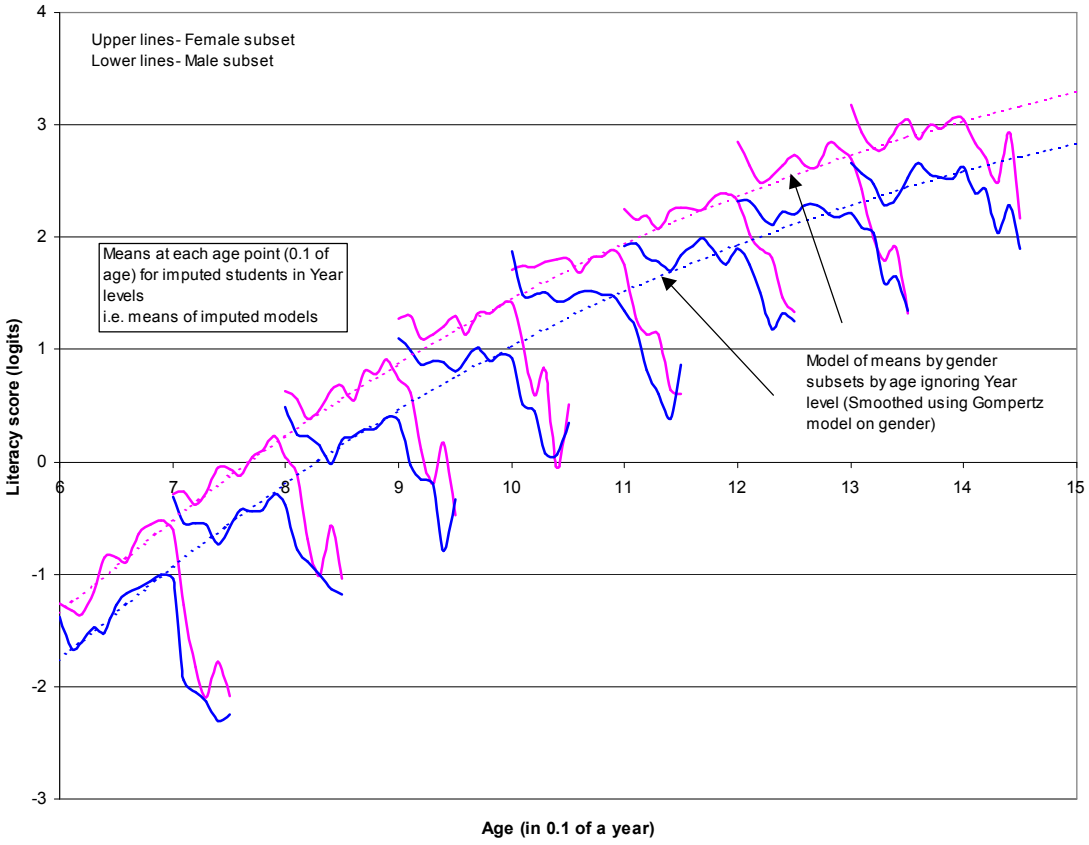
Figure 6.4 Literacy Model by Year level



The within Year level patterns in the model are censored in Figure 6.4: at both ends of a Year level, cases included in Figure 6.3 are censored in Figure 6.4. These censored cases are very small numbers of students with ages well outside the normal age range. The pattern by Year level is generally consistent. There is a very short lead-in for a very small group of younger children (not all shown) who appear to have higher mean scores than those only a month or so older. Then, for the bulk of students, the mean scores increase with age until the highest within normal age range age category for the Year level is reached. Then the set of older students, larger and covering a wider age range than the small group of very young students, produce a tail where mean scores drop off quickly.

Assuming the model approximates reality, it appears that the relative smoothness of the curve in Figure 6.3 is partly a result of the compensatory effects of the older and younger tails, and their proportionately small numbers in the data clusters for each Year level. If there was no age effect the Year level curves would be flat and the composite curve a more pronounced stepwise curve. The data model can be summarised by gender and Year level as shown in Figure 6.5. Unsurprisingly the gender effect is reflected in the model data.

Figure 6.5 Literacy Model by Year level and gender



In the Literacy test model the female mean performance by age is greater than the male mean performance and is consistently represented thus at each year level. The curves of the gender subsets follow the smoothed trajectories shown in Figure 6.5. The smoothed trajectory is obtained by applying a Gompertz model to the gender subsets independent of Year level. The model suggests that the gender effect is consistent over all Year levels, and increasing slightly with increasing Year level/age. The model shows a difference in favour of females is 0.41 logits at age 6, increasing to 0.44 logits by age 12.

The development of the model is based on expecting the mean scores for Year level cohorts to sit on an idealised trajectory. In the South Australian school system around 1997 it is speculated that the patterns identified in Figures 6.3, 6.4 and 6.5 approximate the data for a test assessment for Literacy, applied consistently from Year 1 to 8. This general speculation about Literacy is returned to once the companion story for Numeracy is developed and after the teacher-assessed view of the same learning development is described in Chapter 7.

The trajectory of Numeracy test scores

A similar process as applied for Literacy data is applied in the development of the model for Numeracy data from the 1998 Basic Skills test. The 1998 data are selected to match the

timing of the teacher assessments for Mathematics described in Chapter 7. The data considered in developing the framework are summarised in Table 6.8.

Table 6.8 Numeracy – Mean scores by Year level and Testing Year

	Year Level	3	5	7	9	Growth	Growth	Growth
	Average age	8.6	10.6	12.6	14.5	3 to 5	5 to 7	7 to 9
Cross-sectional	1997	0.05	1.30			1.25		
	1998	**0.13**	**1.34**			1.21		
	1999	0.18	1.27			1.09		
	2000	0.08	1.11			1.03		
	2001	0.13	1.18	2.28		1.05	1.10	
	2002	0.36	1.24	2.46		0.88	1.22	
	2004	0.61	1.46	2.53		0.85	1.07	
	Mean	0.22	1.27	2.42		1.05	1.13	
	1995 (Hungi adjusted)	0.30	1.21			0.91		
	1996 (Hungi adjusted)	0.31	1.24			0.93		
	1997 (Hungi adjusted)	0.21	1.31			1.10		
	1998 (Hungi adjusted)	**0.22**	**1.34**			1.12		
	1999 (Hungi adjusted)	0.20	1.36			1.16		
	2000 (Hungi adjusted)	0.15	1.24			1.09		
	Mean	0.23	1.28			1.05		
Longitudinal	1997 Cohort wave	0.05	1.27	2.28		1.22	1.01	
	1998 Cohort wave	0.13	1.11	2.46		0.98	1.35	
	1999 Cohort wave	0.18	1.18			1.00		
	2000 Cohort wave	0.08	1.24	2.53		1.16	1.29	
	Mean	0.11	1.20	2.42		1.09	1.22	
	1995 Cohort wave (Hungi)	0.30	1.31			1.01		
	1996 Cohort wave (Hungi)	0.31	1.34			1.03		
	1997 Cohort wave (Hungi)	0.21	1.36			1.15		
	1998 Cohort wave (Hungi)	**0.22**	**1.29**			1.07		
	Mean	0.26	1.33			1.07		
NAPLAN	NAPLAN 2008 Numeracy (estimated logits)	0.13	1.16	**2.24**	**2.74**	1.00	1.10	0.50
	NAPLAN SA 2008 Numeracy (reported scores)	*(388.8)*	*(460.4)*	*(536.2)*	*(571.1)*	*(71.6)*	*(75.8)*	*(34.9)*

* Bold values signify values used in estimation of Gompertz model parameters.

Setting the Framework for the Numeracy model

The 1998 views of Numeracy are shown in bold in Table 6.8. For the framework model growth values are the most critical. In the cross-sectional view the growth from Year 3 to 5 appears to have reduced since 1997. However the Hungi adjustment reduces the spread of the cross-sectional growth. Averaged over all views, the growth over two years from Year 3 to 5 is just over 1 logit, slightly less than the general growth for Literacy between these Year levels.

A key difference, relative to Literacy, is the growth in the next two-year period, Year 5 to Year 7. The growth rate for Literacy diminishes, while for Numeracy the growth rate is almost identical to that for Year 3 to 5, based on the cross-sectional, Hungi adjusted and NAPLAN data. Effectively growth in Numeracy learning is linear with age from Year 3 to

Year 7. Based on one point only (the NAPLAN data for SA), the growth rate in Numeracy appears to reduce from Year 7 to 9.

A curve is fitted to the points using a Gompertz iterated solution for the highlighted data for Years 3 to 9, at the average age for each Year level. The process is the same as for the Literacy framework model. Data points are increased in values by 10 logits to avoid any negative values, the Gompertz curve is fitted, and then the resultant curve is adjusted back to the original logit scale. The values of the parameters for the fitted Gompertz curve are a = 15.33, b = 0.37, c = 0.14. The estimated score value at any age value is converted to the original logit scale by subtracting 10. The asymptote (a) of the group mean is effectively 5.3 logits on the original scale as age moves above 20.

Figure 6.6 illustrates the wider spread of values in the cross-sectional view relative to the longitudinal view in Figure 6.7. The most different cross-sectional set is 2004, consistent with Literacy data. The 2004 tests and scaling were provided through a different contractor.

Figure 6.6 Numeracy mean scores-Cross-sectional view with model trajectory

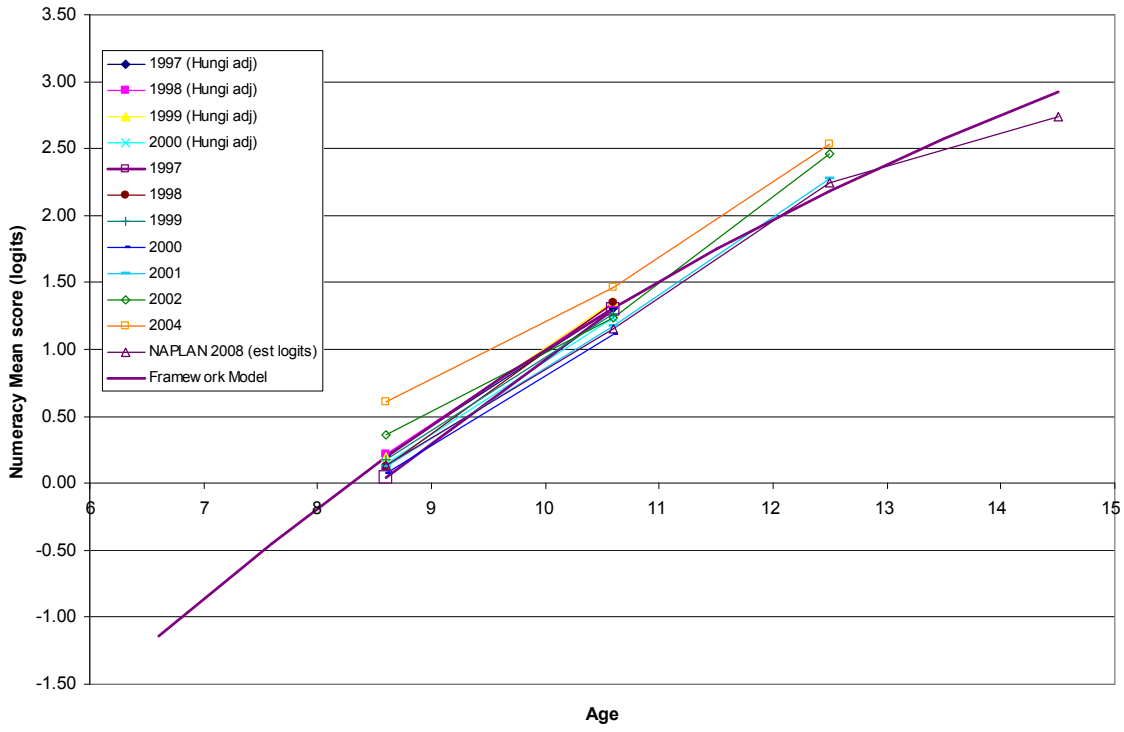


Figure 6.7 Numeracy mean scores –Longitudinal view with model trajectory

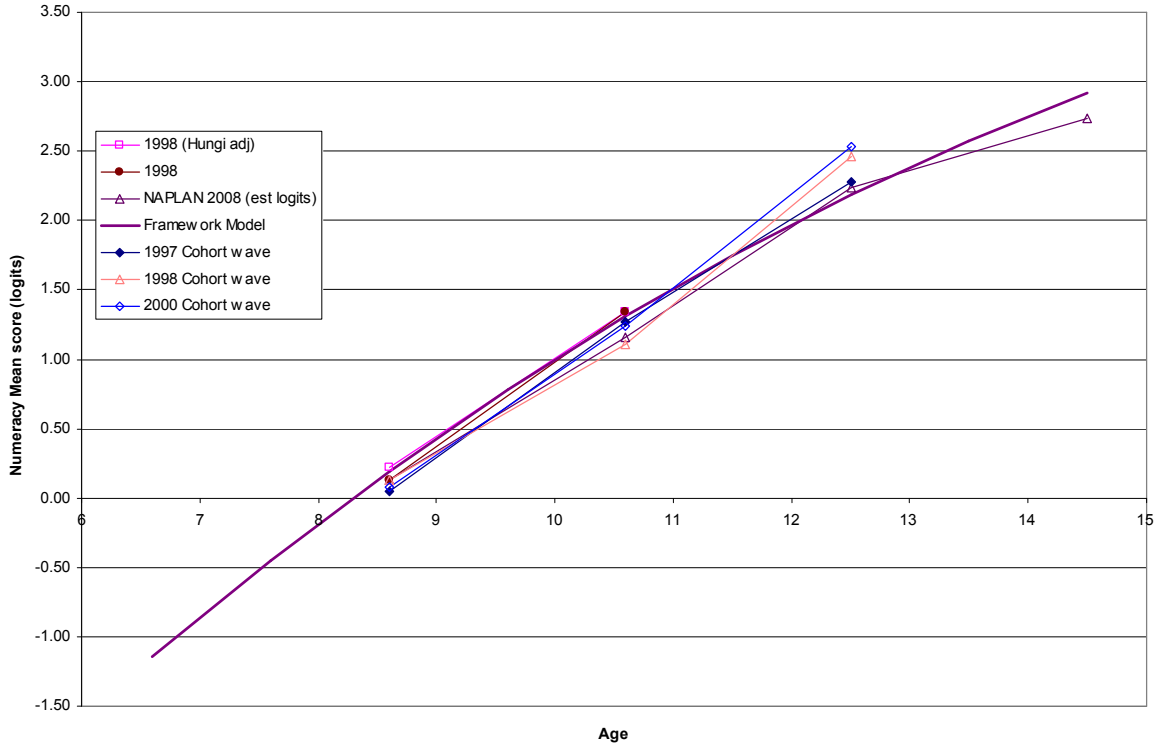


Figure 6.7 shows that the trajectories for most cases in this view are parallel. Consistent with the Literacy data, large differences occur in the means for the Year 7 data in 2002 relative to 2001. The model sets a framework for imputing points. The trajectory at Year 9 (age 14.5) for Numeracy may be poorly modelled. At Year 7 (age 12.6) the fit of the estimated NAPALN data to the trajectory is close.

Adding multiple points to the Numeracy model

As for Literacy, dates of birth were found through matching student records for as many cases as possible. Table 6.9 shows that there were slightly higher percentages of students for whom dates of birth were found than for the 1997 cases for Literacy. Almost 75% of Year 3 and 80% of Year 5 were assigned dates of birth. Ages were then calculated in the following age categories: actual age at testing, conventional age at testing, age categorised into 0.5, 0.2 and 0.1 categories of age at testing. The general statistical characteristics of the sample with birth dates are compared with the original 1998 cohorts in Table 6.9. Apart from the mean for Year 3, most characteristics are very similar.

Via the model framework a process identical to that for Literacy was used to add data points for students. The framework targets for the means of each Year level are shown in Table 6.10. The model was built from the 9567 and 10008 records for Years 3 and 5 respectively. Independent samples of 7990 were taken for Years 1 to 3 and 5. Original Year 3 and 5 cases

were sampled and then combined to create 7990 cases for Year 4. Year 7 files for 2001 and 2002, as used for Literacy, were used to create cases for Year 7, Year 8, and blended with Year 5 to create cases for Year 6.

Table 6.9 Numeracy-comparison of original records with subsets assigned dates of birth

	Sample with Birth Dates		Full cohort Statistics	
	Year 3	**Year 5**	**Year 3**	**Year 5**
Mean	0.20	1.38	0.13	1.34
Median	0.22	1.38	0.22	1.38
SD	1.32	1.12	1.36	1.16
Skewness	-0.09	-0.25	-0.15	-0.42
Kurtosis	4.39	5.85	4.54	6.38
IQR	1.71	1.35	1.71	1.35
N	9567	10008	12794	12471
% with DOB			74.78%	80.25%

A statistical summary of the match of the Numeracy model to the Framework model is documented in Table 6.10. The scores for each student were systematically adjusted to bring the mean for the year level as close as possible to the target. The spread characteristics reflect the original data sources. An anomaly in Year 7 data mentioned in footnote 25 also applied for Numeracy but for a separate set of students who had Literacy scores in the expected range but no Numeracy score. The anomaly was discovered after the model had been developed. As a result the cases omitted in the original files were deleted from the final model and the remaining records adjusted to match the Target means. This caused a small loss of records in the Year 6, 7 and 8 models, reflected in the count of points in these Year levels being less than the 7990 target.

Table 6.10 Numeracy Model-main statistical characteristics

Numeracy Model by Year Level	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Model (Targets for each YL)	-1.16	-0.45	*0.20*	0.79	*1.32*	1.80	2.19	2.57
Means of actual or imputed points	-1.16	-0.45	*0.19*	0.78	*1.31*	1.75	2.22	2.59
Medians of actual or imputed points	-1.16	-0.40	*0.20*	0.82	*1.32*	1.72	2.18	2.54
SDs of actual or imputed points	1.35	1.34	*1.32*	1.22	*1.12*	1.13	1.12	1.10
Skewness of actual or imputed points	-0.11	-0.06	*-0.03*	-0.01	*-0.28*	0.04	0.32	0.21
Kurtosis of actual or imputed points	4.34	4.38	*4.19*	4.51	*5.70*	4.93	4.10	3.80
IQR of actual or imputed points	1.74	1.66	*1.71*	1.51	*1.35*	1.37	1.44	1.44
Count of actual or imputed points	7989	7990	*7900*	7990	*7990*	7916	7851	7872

The model compared to the Frameworks- Numeracy

Figure 6.8 compares the model of individual student scores with the target framework. The model follows the target trajectory well. The path of the trajectory is similar to the Literacy equivalent (Figure 6.3). The target points sit on the curves of the means. The intermediate points wobble along the general trajectories as should be expected from the stochastic nature

of the learning process and the known within-Year level age patterns. As the points at Year 3 (age 8.6) and above are derived from actual data, the model is assumed to approximately match the distribution of scores that would apply if all students in the model had been tested. Below Year 3 the extent to which the model matches reality relates to the steepness of the trajectory in these Years. To achieve the match to the Gompertz determined trajectories, data points were spread more widely by age.

Figure 6.8 Comparison of Numeracy Model with the Framework Model

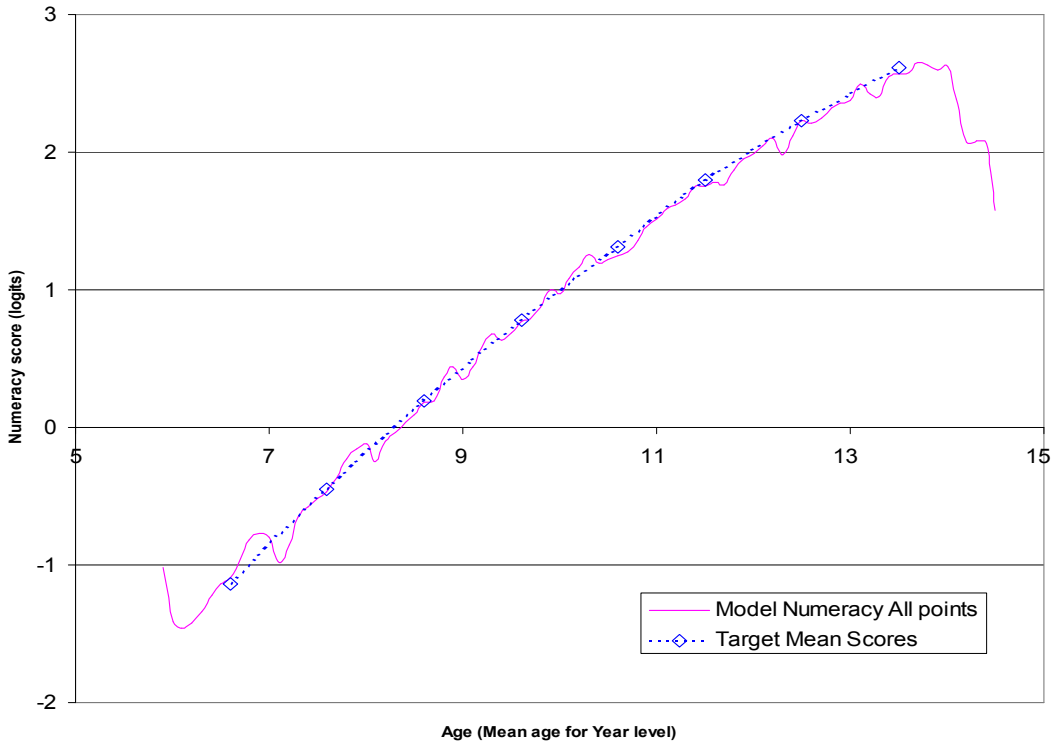
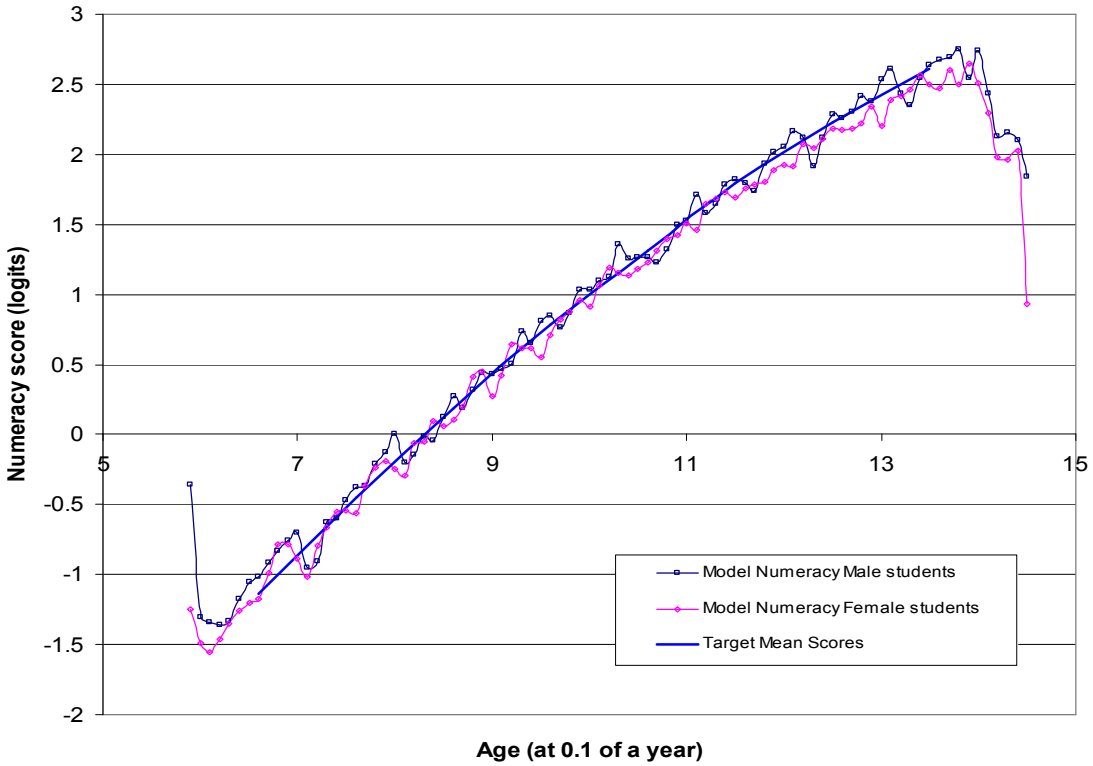


Figure 6.9 compares the trajectories of the mean points for each 0.1 of age, by gender. Recent evidence from NAPLAN (2008) indicates slightly higher scores for males apply, increasing with age (estimated to be 0.1 logits at Year 3 and 0.2 logits at Year 9). Data for 1998 indicate small differences of similar amounts (0.07 logits in favour of males at Year 3, 0.09 at Year 5). These differences contrast with the larger score advantage for females in Literacy at all levels, starting at about 0.3 logits and increasing with age.

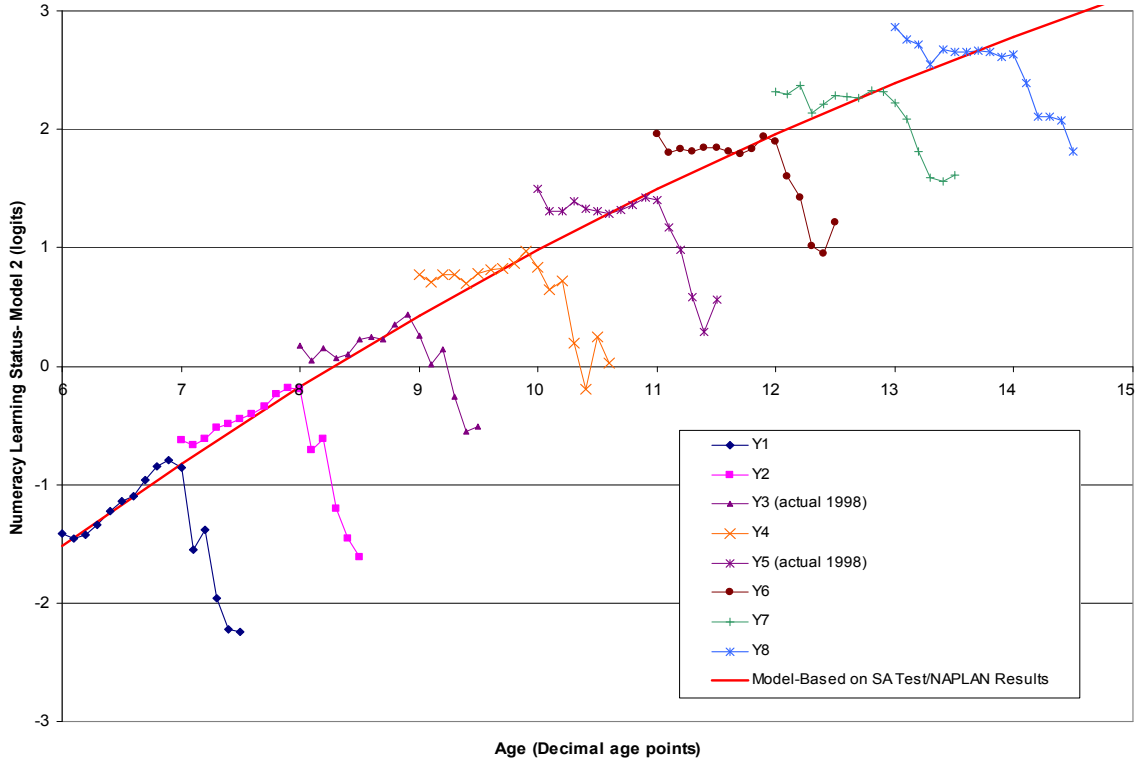
Consistent with the small advantage to males in Numeracy, the model generates a summary for males that tends, on average, to be greater than the female summary as shown in Figure 6.9. A more refined gender analysis from the model is covered in subsequent sections.

Figure 6.9 Numeracy Model by gender



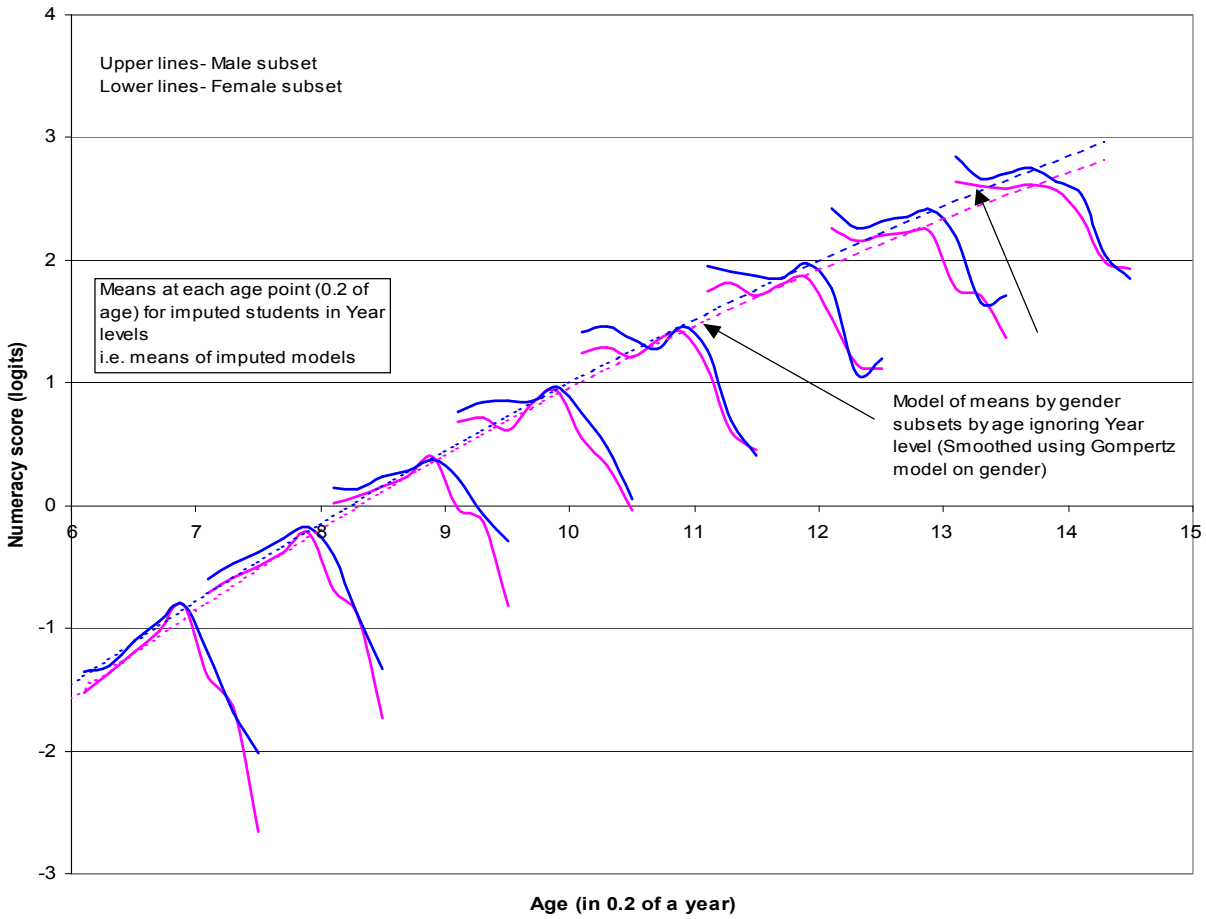
When the model is summarised by Year level (Figure 6.10) a very similar pattern is obtained for Numeracy as is found for Literacy. At lower Year levels the gradient with age within a year level appears to be greater (although for Years 1 and 2 the effect is artificially created by the model development process). At all Year levels, students older than the normal age range for the Year level have lower mean scores and generate a tail of diminishing scores. This is consistent with the examples reported in Chapter 5 where data summaries from a wide range of test samples show this specific pattern of learning status by age within a Year level. Also consistent with Chapter 5 the gradient of the effect diminishes with increasing Year level.

Figure 6.10 Numeracy Model by Year level



Year level data in the model can be disaggregated by gender as shown in Figure 6.11. The summary in this case uses age categories of 0.2 age divisions to smooth the variability shown in Figure 6.10. Consistent with the general understanding of performance by gender, the male performance in Figure 6.11 is marginally higher than that of females at each Year level, with the difference growing with age. The trajectory for each gender group can be obtained by fitting the Gompertz expression separately. The gender trajectories start close together with a slight male advantage as age (Year level) increases. This is consistent with the general pattern in the NAPLAN (2008) data.

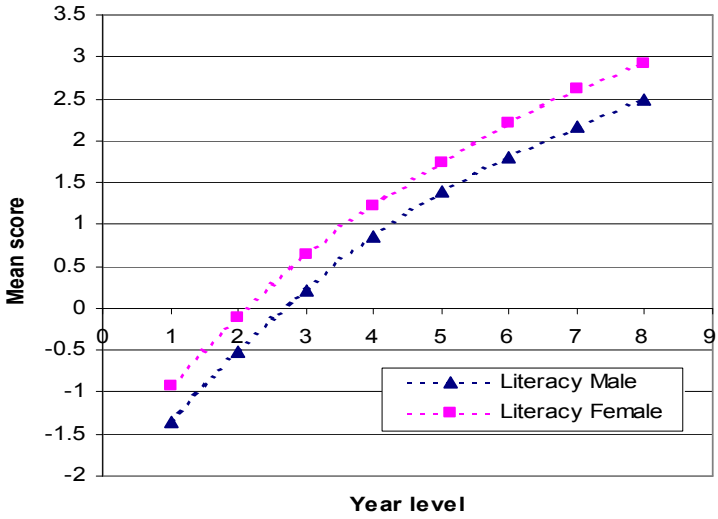
Figure 6.11 Numeracy Model by Year level and gender

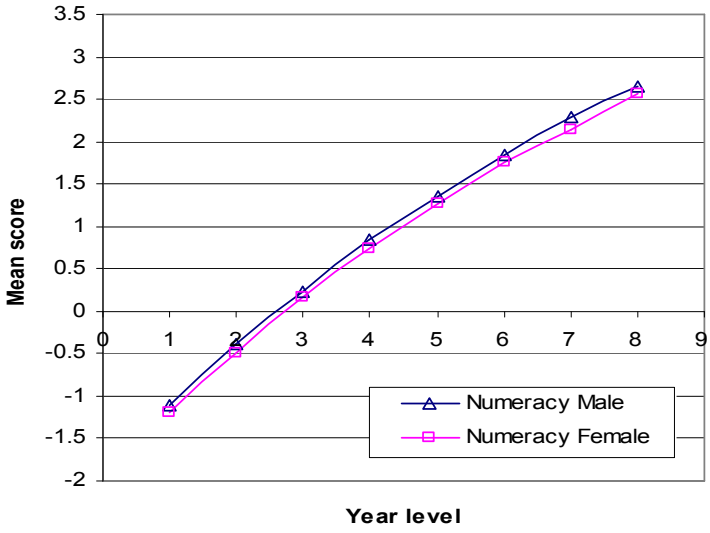


The models above were developed for two calendar years, for Literacy in 1997 and Numeracy in 1998, to estimate data for the non-tested Year levels and to understand the age effects within and across Year levels when consecutive cohorts of students are tested.

The panels below show Year level views illustrating the benefit of the models being developed on an individual student basis. Summaries can be made in a number of ways. In Figure 6.12 the Year level view is provided. The panels illustrate that the general models, developed without consideration of gender, enable a mean score to be estimated for gender by Year level or age.

Figure 6.12 Summary of the Literacy Model and Numeracy Model by Year level and gender





The gender views that come out of the models are consistent with those from other sources. Based on the UK Statistical First Release 19/2009 (2009, Table 5) for example, the trend by gender at Key Stage 2 (11 year olds) for mathematics shows the male average point score for 2009 at 27.7 points compared with 27.4 points for females (0.3 point difference), which has persisted for a number of years. On the other hand for English language there is a difference in favour of females of 1.6 points (28.1 for females, 26.5 for males). The specific patterns for both learning areas by gender have persisted for a number of years, at least since 2004 (Statistical First Release 19/2009, 2009). Accordingly, the models developed for the South Australian test data for Numeracy and Literacy follow, in general terms, the trends found elsewhere.

By building the model at a student level, a richer summary of the general patterns of performance by gender and age has been developed. For example the patterns by age within a

Year level can be shown. The models are based on an iterative fitting of a curve through four Year level score means (Years 3, 5, 7 and 9) using a Gompertz model. Other curve-fitting processes may produce equivalent results. The assumption of predictable growth in mean learning per Year level and by age is critical to the models. The evidence in Tables 6.4 and 6.8 indicates that average rates of learning growth with age have been approximately consistent across calendar years and where variations occur they can often be explained by test calibration variability. Assumptions about specific rates of growth with age and Year level would be unnecessary were assessment data available for all Year levels.

For two Year levels (3 and 5) the models use a large sample from the full cohort test data for the appropriate collection years. For Year 7 actual data are used, but are a blend of two collection years 5 years later. An estimated Year 9 mean influences the trajectory of the data.

Years 1 and 2 data points generated for the models are the most artificial since they are a transformation of individual scores from Year 3 to make cohort means match the framework. To achieve this a spreading of the data to increase the within Year level age gradient is required. The actual trajectory for the lower years is unknowable by the usual pencil and paper testing processes. The extrapolation from the known points, while plausible to the author at least, and generally consistent with trends from other sources as discussed in Chapter 5, is highly speculative. However the gradient of the lower trajectory is conservative relative to some estimates of the rates of learning at lower years (Hill, Bloom, Rebeck, Black & Lipsey, 2007 discussed in Chapter 5).

Summary

The purpose of the chapter was to report actual test data for Years 3 and 5 for South Australia and establish the quality of these test assessments. These data covering only two Year levels were then extended using the general findings for the trajectories of growth of learning status for cross-sectional groups (established to approximately match longitudinal groups) to develop a framework for the trajectory of the means at all Year levels. These framework trajectories were developed for literacy and numeracy.

Samples of student records taken from the actual Year 3 and Year 5 data for the appropriate calendar years (1997 and 1998), supplemented by Year 7 student records for 2001 and 2002, were then blended and means re-centred to fit the framework trajectories. Year 1 and Year 2 samples were stretched to match the framework trajectories and to match the general shape required by the models developed in Chapter 5.

The general data developed were then summarised from age, Year level and gender perspectives to report an estimated but speculative view of what summaries of learning status

of a sample of students tested at all Year levels form 1 to 8 might look like. These summaries provide one basis for comparing teacher judgement assessments of the same cohorts to test assessments.

In the next chapter the same general learning areas are assessed but based on teacher judgement assessments rather than tests. How the two approaches compare is addressed in Chapter 8.

Chapter 7: South Australian teacher judgement assessments: 1997 and 1998

> …profiles function as a framework for assessment and reporting and do not in themselves constitute an assessment method. What they do allow, however, is for teachers, schools and school systems to communicate about student progress and achievement using a language and standards which are consistent across classrooms, schools and school systems.
>
> Hill, 1994, p. 38.

This chapter summarises the key findings from the South Australian teacher judgement assessments using the Statements and Profiles for Australian Schools (SPFAS) approach. The history and detail of the approach are described in Chapter 3. The data collection processes and the assessment role of teachers are summarised here briefly.

Teacher judgement assessments are addressed in their own right without reference to checks against alternative assessments. This is done to appreciate the capacity of teachers to make assessments of individual students on an eclectic loosely specified basis but against general criteria (SPFAS) provided as a map for the development of learning in each strand considered. An overview summary of the assessment results is presented in the adopted metric of those assessments: profile or level units.

The overview then leads to the important question, "How well do the South Australian teacher assessments match test assessments?" To consider that question adequately, the assessment scales need to be converted to a metric common with that of the tests. Processes for converting teacher assessments to the test metric and then comparing them with each other are addressed in the next chapter.

The data collection revisited

Detail of the data collection process is covered in Chapter 3. In brief, the data collections of 1997 and 1998 were identical in most respects. While in 1997 four learning areas were included, only the English learning area data are used in this study. The teacher judgement assessment survey was conducted in October; this timing has a small impact to be considered when test and teacher assessments are compared, as the tests had been conducted in August, two months earlier. In 1998 the remaining four learning areas were included. Of these, the Mathematics learning area is the focus for this study. The teacher judgement assessment survey was conducted in the same month as the tests (for Years 3 and 5) making the timing of the assessments of no further concern in the comparison of data.

Survey software was used to manage the random selection of students and the learning area to be reported for each student. As a result teachers did not know in advance which students

they would be required to report on, nor on which learning area. Each teacher reported for five randomly selected students from that class, by indicating which of the eight described levels for a strand in the learning area had most recently been achieved by each student. The teacher then reported on each student's progress towards achieving the criteria for the next level by clicking on a continuous line. This line was segmented but this was not indicated to the teachers. A click on the line activated one of nine segments, leading to an indication of progress in 0.1 segments. Data files for the collections included about 120,000 records in 1997 and over 200,000 in 1998. Each record was one student/strand/rating event. The unique identifier for the student was required to manage the screen presentation to the teacher and was preserved in the collection process to add other identifiers to the file (gender, socio-economic status of the family, language background, and indigenous status) found from the general statistical records of the education department.

Rothman (1998, 1999) analysed strands within learning areas as separate summaries for each strand. In the current analysis records were restructured so that individual strand assessments (for English and Mathematics) were consolidated for each student. The purpose of the record restructure was to allow a mean assessment in the learning area to be made for each student, consistent with the general principle of test design where composite strands are combined in the test design and a general overall score for each student calculated. In the cases of a test the strand equivalent data (e.g., reading and writing in English) can be analysed separately and individual item performance for each student investigated. In the teacher judgement assessment data drilling down below the strand is not possible.

Once the files were structured as consolidated student records it was possible to attach the gender and date of birth to calculate the ages at assessment for each student. For 1997 the restructuring of records resulted in 7871 student cases over 8 Year levels, approximately 1000 cases per Year level, and about 100 cases per age categorised at 0.1 of age (that is about 100 cases for each month of age). On the basis of 5 students assessed per teacher, the data represent the assessments of 1500 teachers. For 1998, 12050 student cases over 8 Year levels were assembled, approximately 1500 cases per Year level, and about 150 cases per age categorised at 0.1 of age. On the basis of 5 students assessed per teacher, the data represent the assessments of 2400 teachers.

The data collection has a number of inherent independent replications; collection period (1997, 1998), different learning areas (Literacy and Numeracy), eight Year levels for each collection, primary versus secondary teachers, and includes 20000 students and almost 4000 teachers overall.

The English Learning Area

The general statistical characteristics of the data for English are listed in Table 7.1. The average ages at assessment differ consistently by one year of increasing age for each increase in Year level. The average age is greater by 0.2 of a year of age than for the Mathematics data reported in Table 7.2 and the data and models developed in Chapter 6, due to the two-month difference in assessment periods. The data are averages of two judgements (Reading and Writing) rather than for each strand separately.

Spread of assessments and scale use

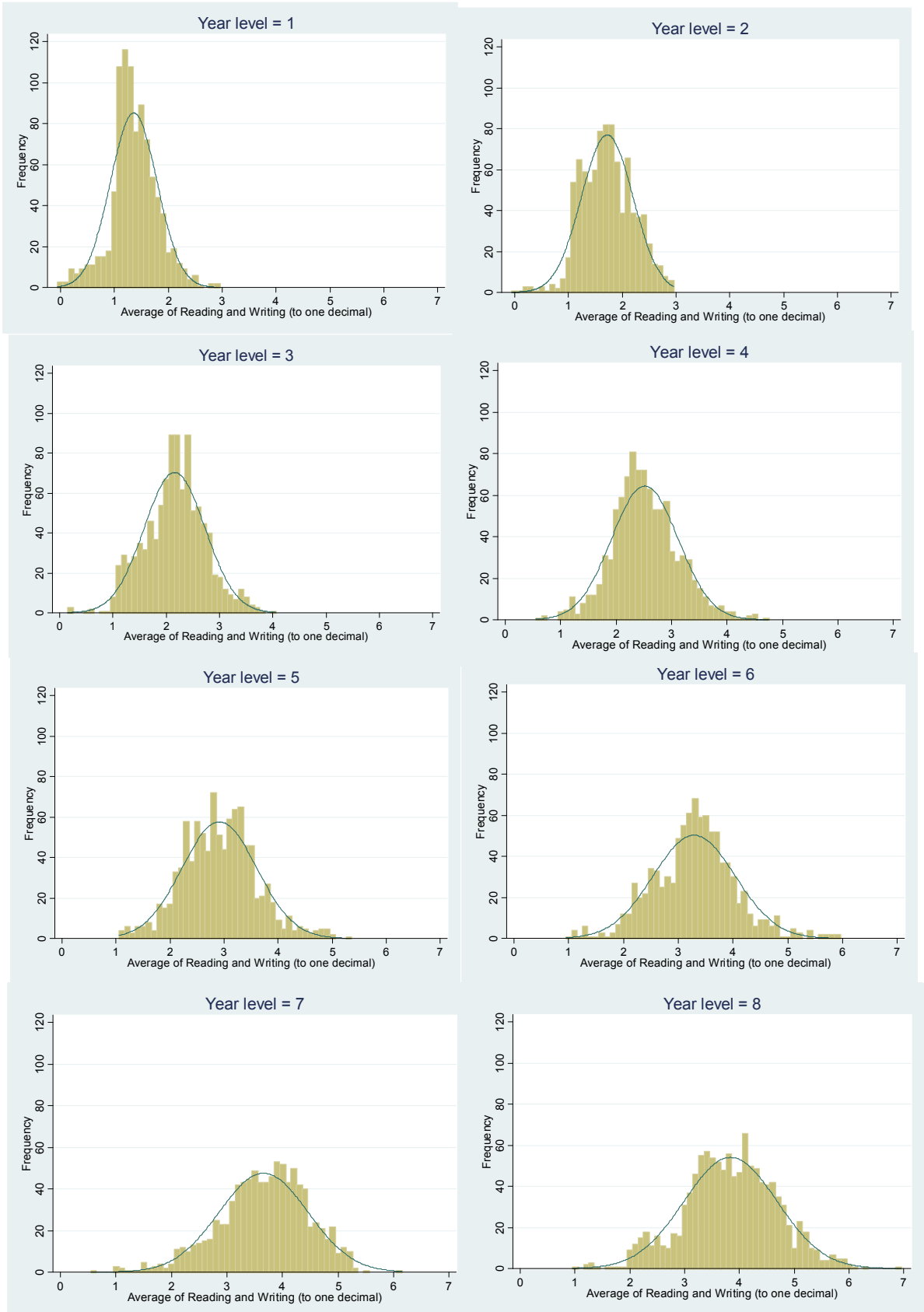
Issues arise from the response format and the history of the design of the teacher assessment collection. Did teachers use the full range of the scale, based on the developmental range of the students for a given Year level? What form does the distribution of the assessment results take?

A view of the spread of the teacher assessments along the assessment dimension, and the likely use of the full spectrum of response possibilities, is obtained in the panels in Figure 7.1. The histograms indicate that the full spectrum of responses appears to have been used by teachers in their assessments. In general terms, the averaged English assessments, based on equal weighting of the Reading and Writing, are spread around the mean and fit the shape, for most assessment values, of the superimposed normal curve. There are exceptions.

For Year 1, scale positions just above 1 are very well used. These are points that indicate that the student has met the criteria for level 1 but has not progressed much further. While these points appear over represented and as a consequence some other points under represented (just below 1 as examples), the panel shows that the full range of assessment points are used. Similar over and under representation are shown in other panels.

For Year 2, a point just past level 2 stands out due to points missing either side, although the segment itself sits close to the super-imposed curve. For Year 3 the early points on the scale from 2 to 3 are over represented. A similar effect is observed for Year 4. At Year 5 the effect has moved to the beginning of level 3 and remains in this segment of the scale for Year 6. For Years 7 and 8 the first segment from 4.0 to 4.5 is over represented relative to a normal distribution. The effect is also obvious for the beginning of 3.0 to 3.5 for Year 8. Year 8 is the first year of secondary school and thus reflects the assessments of secondary teachers as against those of primary school teachers. The data confirm that teachers used the full range of points on the (hidden) underlying 10-point scale (unwittingly since they responded to a line rather than assigning numbers) and did so with a preponderance towards the early segments of each new level.

Figure 7.1 English 1997 – Histograms of score distributions by Year level



Learning status trends with Year level

The means and medians, as shown in Table 7.1 are close, usually differing in the second decimal place only, indicating that the cases are approximately evenly distributed around the mean for each Year level. This is further illustrated in Figures 7.2 and 7.3 where the relationship of each English strand and the combined Reading and Writing strands with Year level/age is shown to be linear with Year level up to Year 7. Year 8 is an exception. This contrasts with the general shape of the mean test scores with Year level/age described in Chapters 5 and 6 where the means of IRT based measures show decelerating growth with Year level/age: IRT measures are not linear with Year level or age.

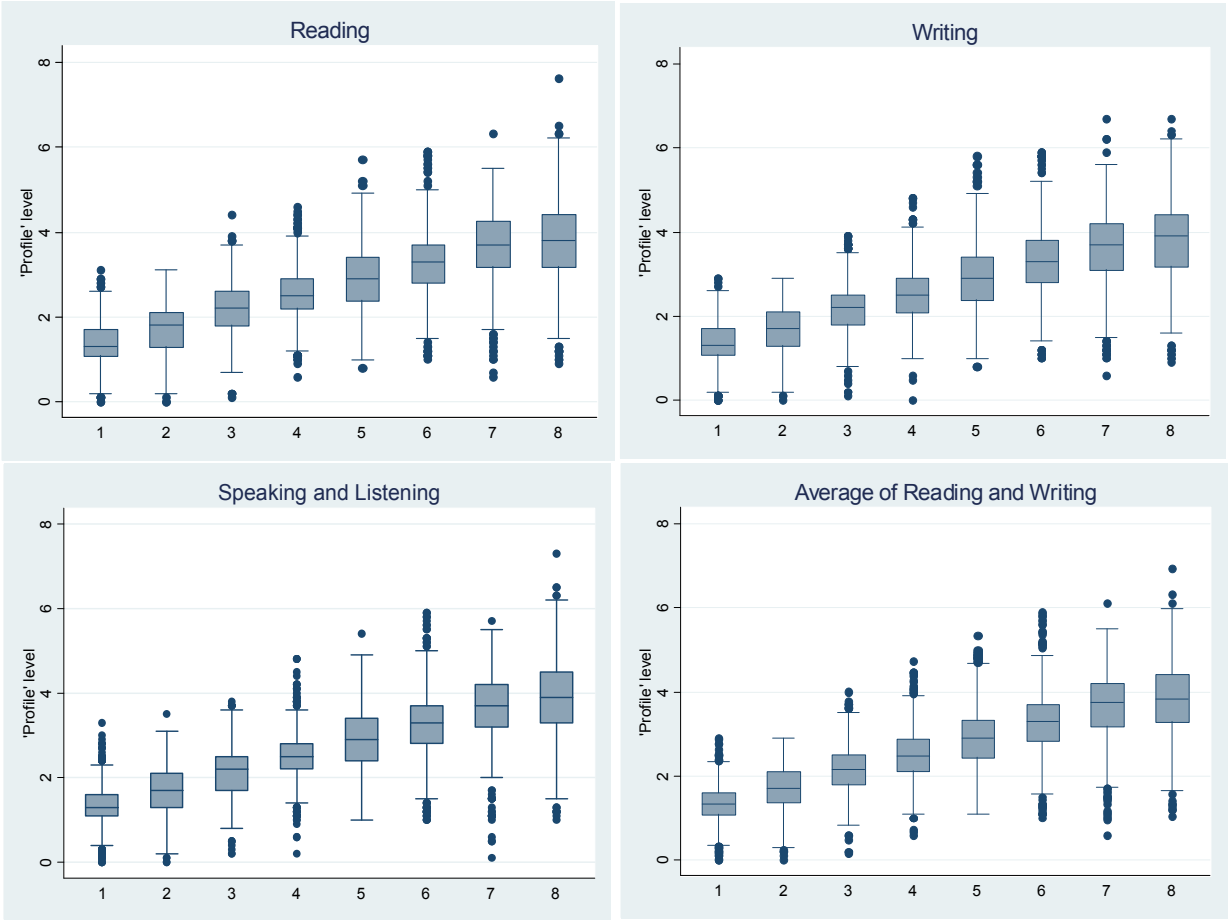
Table 7.1 English Learning Area by Year level–1997: General Statistics

		Average of Reading and Writing values (in Profile level units)							
Year level	**Average age at assessment (October 97)**	**Mean**	**Median**	**SD**	**IQR**	**SE of mean**	**Skew-ness**	**Kurtosis**	**N**
1	6.81	1.36	1.33	0.43	0.50	0.01	-0.04	4.05	923
2	7.79	1.73	1.70	0.48	0.73	0.02	-0.04	3.06	926
3	8.79	2.16	2.17	0.57	0.70	0.02	-0.03	3.23	1005
4	9.78	2.52	2.47	0.60	0.73	0.02	0.25	3.61	969
5	10.79	2.91	2.90	0.69	0.90	0.02	0.22	3.37	996
6	11.78	3.29	3.30	0.75	0.87	0.02	0.15	3.88	945
7	12.81	3.66	3.73	0.80	1.00	0.03	-0.49	3.39	956
8	13.79	3.84	3.83	0.85	1.10	0.03	-0.10	3.17	1151
All	**10.39**	**2.72**	**2.63**	**1.08**	**1.60**	**0.01**	**0.28**	**2.52**	**7871**

The standard deviation (SD) and the inter-quartile range (IQR) increase with Year level and age. This phenomenon is consistent with the Rowe and Hill (1996) observations for teacher judgements assessments in Victoria and consistent with the general linear relationship with grade, often indicative of a grade equivalent rescaling (Schulz & Nicewander, 1997). This key observation will be discussed in more detail following the description of the mathematics teacher judgement assessments. It offers a possible understanding for how teacher judgment assessments are made and why their distributions differ from those of IRT test assessments.

Figure 7.2 shows that the linear relationship of the medians for each Year level is consistent across strands.

Figure 7.2 Teacher Judgement assessments - English Learning Area 1997 by strand and Year level



The relationship of the mean teacher assessment with Year level is linear up to Year 7 as shown in Figure 7.3. The median scores vary around the linear trajectory of the means. Both the SDs and IQRs are shown to increase with Year level. This is illustrated again in Figure 7.4 where data points are plotted in 0.1 of a year of age, rather than at the average age of the Year level group. The points representing the mean of all the students in all the age categories of 0.1 of an age follow a linear trajectory with age, on average, with only a few points deviating from the general trajectory. A linear regression of the mean assessment scores with age up to Year 7 has a gradient of 0.374 profile level units per year of age. Using the mean of each age grouping eliminates the variance within age. The line of best fit (up to age 13.5) has a very high R^2 (above 0.99) suggesting a very good fit of the line to the means up to Year 7. Year 8 data indicate that secondary teachers report students at a point lower than the previous primary annual improvement would predict.

Figure 7.3 Teacher Judgement assessments - English Learning Area 1997: means, medians, standard deviations and inter-quartile ranges, by Year level

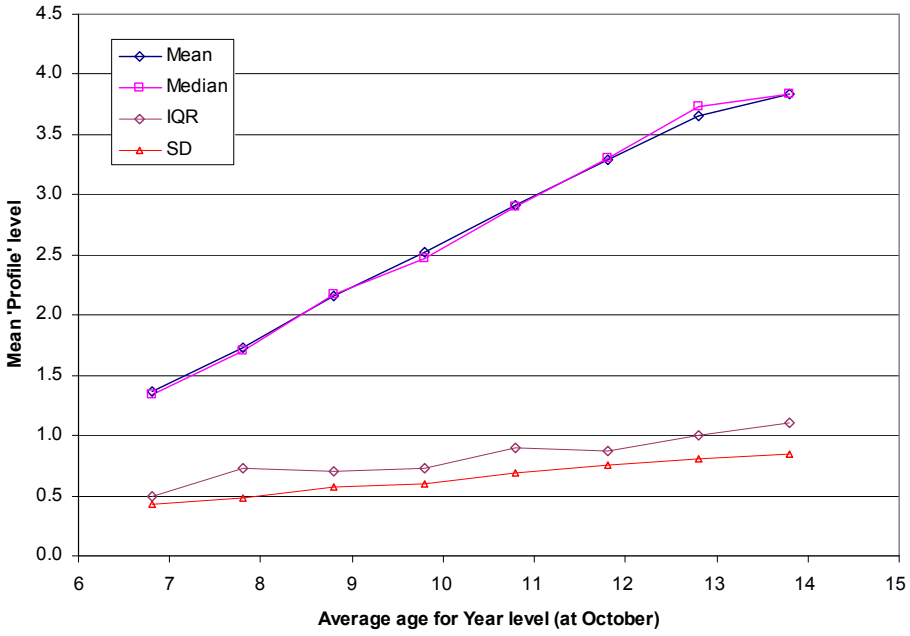
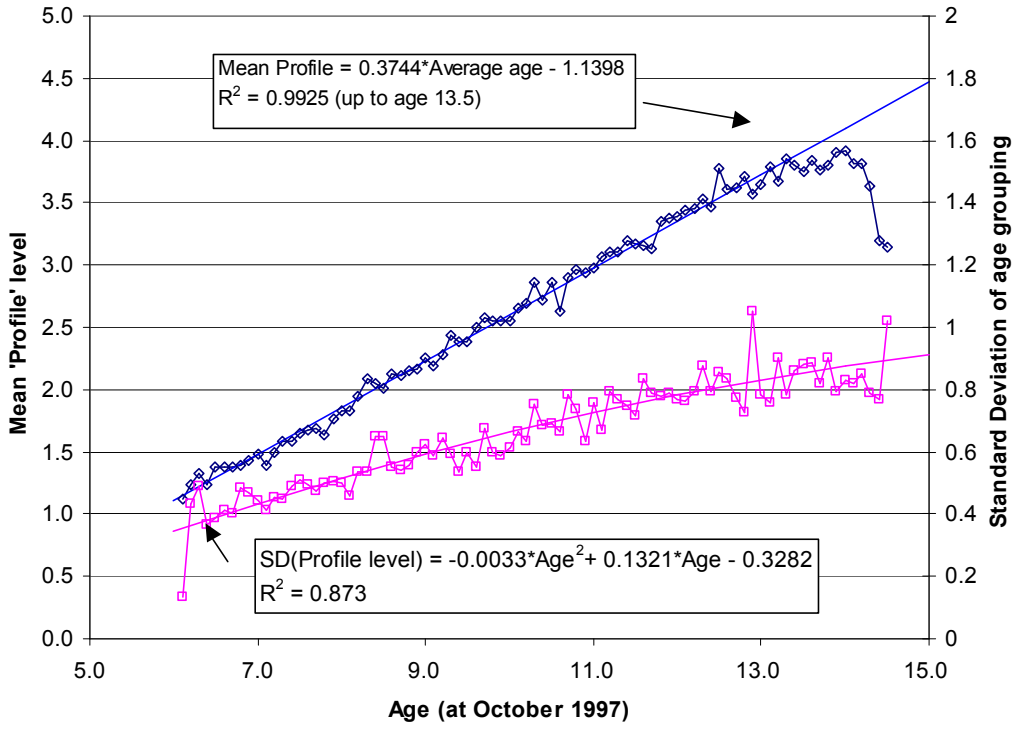


Figure 7.4 Teacher Judgement assessments - English Learning Area 1997: Mean profile level of Reading and Writing strands combined, by age



The increase in the SD with age is essentially linear, with slight levelling out from age 12. This is reflected by a quadratic curve fitting the data points slightly better than a straight-line function. As referenced earlier the phenomenon of linearly increasing means with Year level and age and increasing SDs is consistent with findings for Grade equivalent assessments.

Effect of age at assessment on the relationship of learning with Year level/age

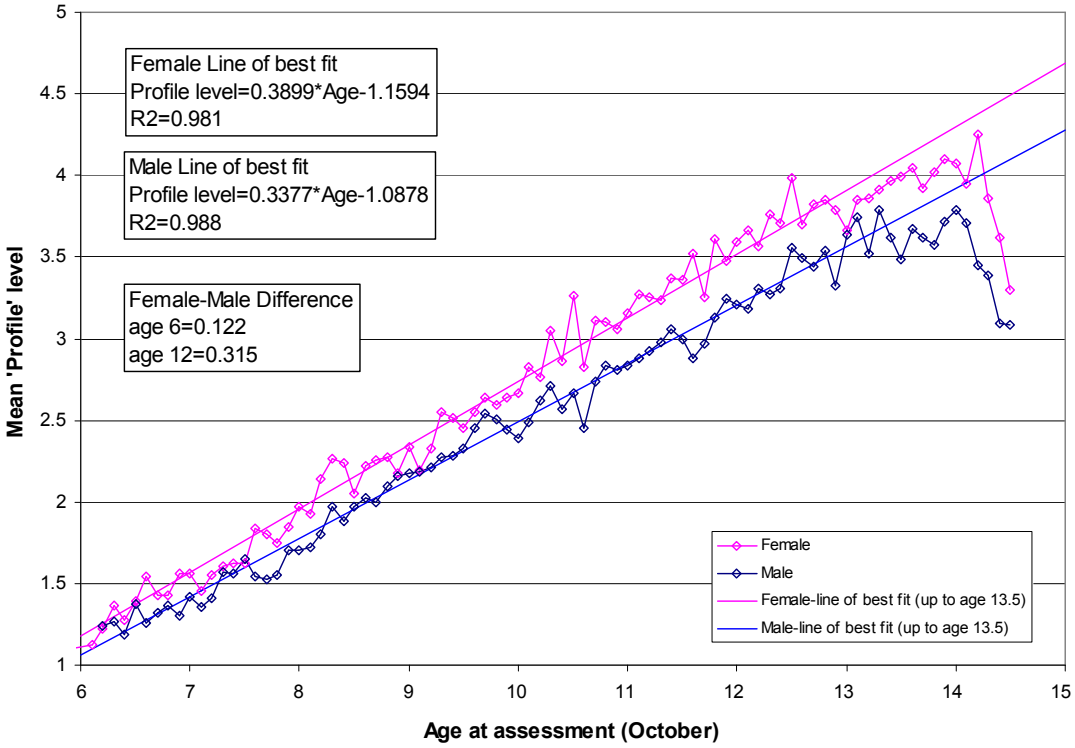
The age effect is shown to apply consistently across age categories of 0.1 of a year. A consistent linear gradient of 0.374 of a profile level unit per year of age from teacher judgement assessments applies up to age 13. The effect of a two to three month difference in the age (time) of assessment can be explored by the use of the regression expression found to fit the age data points in Figure 7.4. The gradient of 0.374 profile level units per year of age is also the growth in the means in successive Year levels assessed at the same time point, since the points are one year apart. A simple estimate of the mean difference between the August assessment date (test) and the October assessment date (teacher judgement) can be based on an assumption of about 2.5 months time difference, equivalent to 0.2 of a year. Based on the linear growth with age expression, a 0.2 difference of a year in age at assessment leads to a 0.07 profile level unit difference. This difference is approaching the pre-set resolution of the teacher judgement scale of 0.1 profile level units.

This analysis suggests that an age adjustment should be considered before the test and teacher assessments are directly compared. For a single comparison as applies in this thesis the age/time of assessment adjustment is less necessary, but not applying it will produce a relationship of test and teacher assessments that is slightly displaced. The issue is considered again in Chapter 8.

Gender differences in the English Learning Area

Figure 7.5 illustrates the difference between teacher judgement assessed learning trajectories by gender. Consistent with assessment summaries using test data, female students have assessments consistently higher than do males for any given age.

Figure 7.5 Teacher Judgement assessments - English Learning Area 1997: Mean profile level of Reading and Writing strands combined, by gender of students

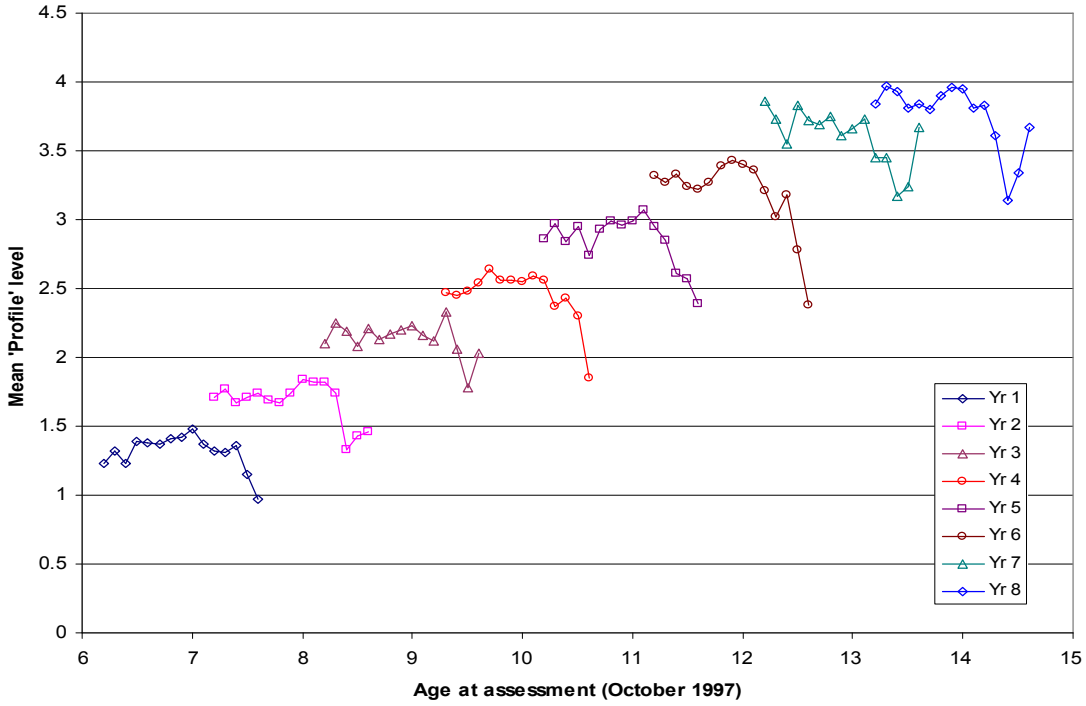


The gender difference based on the teacher assessments appears to increase with age and is reflected in the slopes of the regression lines fitted to the linear segment of the growth from age 6 to age 13.5. The trajectory is clearly different from Year 7 to Year 8 (age 12 to age 13) and is therefore truncated for the line of best fit up to age 13.5, only part way into the data points that represent Year 8. Based on regression lines, the difference at age 6 is 0.122 profile level units. By age 12 the difference is 0.315 units. The general pattern is consistent with that shown in Figure 6.5, where the difference between mean female and male test scores appears to increase with age.

Within Year level trends by age

Figure 7.6 illustrates the trends by age within Year level. At the gross scale of the profile level unit it is difficult to observe how well the trend in test scores with age, shown in Chapters 5 and 6, is reflected in the teacher data. The tails for students beyond the normal age range for the Year level are similar to the pattern for Literacy in Chapter 6. A more detailed comparison is addressed in Chapter 8, once the teacher and test data are brought to an approximately common scale.

Figure 7.6 Teacher Judgement assessments - English Learning Area 1997: Mean profile level of Reading and Writing strands combined, by age within Year level



Since the data were collected in October, the normal range for a Year level is from x.2 years to x+1.2 yrs, where x is the age appropriate to the Year level. The first point for each Year level commences at x.2. The small numbers of cases below this age have been censored for each Year level. For the lower Year levels there appears to be a slight gradient with age until the over-age cases are reached. The pattern is less regular than for the test data and model. There are fewer cases for each age point in the teacher data (i.e. lower *n* relative to the test data) thus potentially larger variation from the general trend pattern at each age point. The within-Year level age effect seems to disappear by Year 7.

The major characteristics of the teacher judgement assessment data for English have similarities with those for Mathematics, based on teacher judgment assessments made one year later.

The Mathematics Learning Area

The general statistical characteristics of the mathematics data are listed in Table 7.1. The mean ages at assessment increases consistently by one year of age for each increase in Year level. The mean age at each Year level is lower by 0.2 of a year of age than the English data reported in Table 7.1 as the assessments all occurred in August 1998. The means for each Year level are consistent with the test data and models developed in Chapter 6.

Spread of assessments and scale use

As for English the first aspect of the data is the distribution of the assessments and the extent to which teachers used the full range of assessment points available to them. The spreads of teacher judgement assessments along the assessment dimension (Figure 7.7) are similar to those for English, and indicate the use of the full spectrum of response possibilities. The data are the means of five judgements per student in each of the mathematics strands rather than for each strand separately (see Figure 7.8 for the individual strands). The histograms indicate that the full spectrum of response possibilities appears to have been used by teachers in their assessments. The mathematics assessments are spread around the mean and fit the shape of the superimposed normal curve for most assessment scale values. As for English there are exceptions. At Year 1 the scale positions just above 1 are very well used. These points indicate that the student has met the criteria for level 1 but has not progressed much further. While these points are over represented, from a normal distribution perspective, some other points are under represented (e.g., just below 1), the panel shows that the full range of assessment points are used. Similar over and under representations are shown in other panels.

On the assumption that the assessments should be normally distributed it would seem that teachers may under report students who have yet to reach the criteria for level 1 (Years 1, 2, 3), level 2 (Years 4, 5, 6), level 3 (Years 6 and 7) or level 4 (Year 8). The distributions for each year level appear to have regions of missing values to the left of the profile boundaries.

The SDs and IQRs increase with increasing year level, as for English, and are shown in Table 7.2.

Learning status trends with Year level

The trends with Year level for each strand of mathematic are shown individually in Figure 7.8. All show the same general pattern of linear growth with Year level up to Year 7 and then less growth in Year 8. The data for each student for each strand are combined into a grand average, shown in the lowest right panel. This box-plot averages the variations of the assessment in each strand for each student to an average assessment value, similar to the process that applies in the Numeracy total test score. The average of all strands is plotted by Year level (placed at average age for the Year level) in Figure 7.9 and tabulated in Table 7.2. The trajectory is linear up to Year 7. The medians and means are very close indicating that the distributions are reasonably well centred on the mean.

Figure 7.7 Mathematics 1998 – Histograms of score distributions by Year level

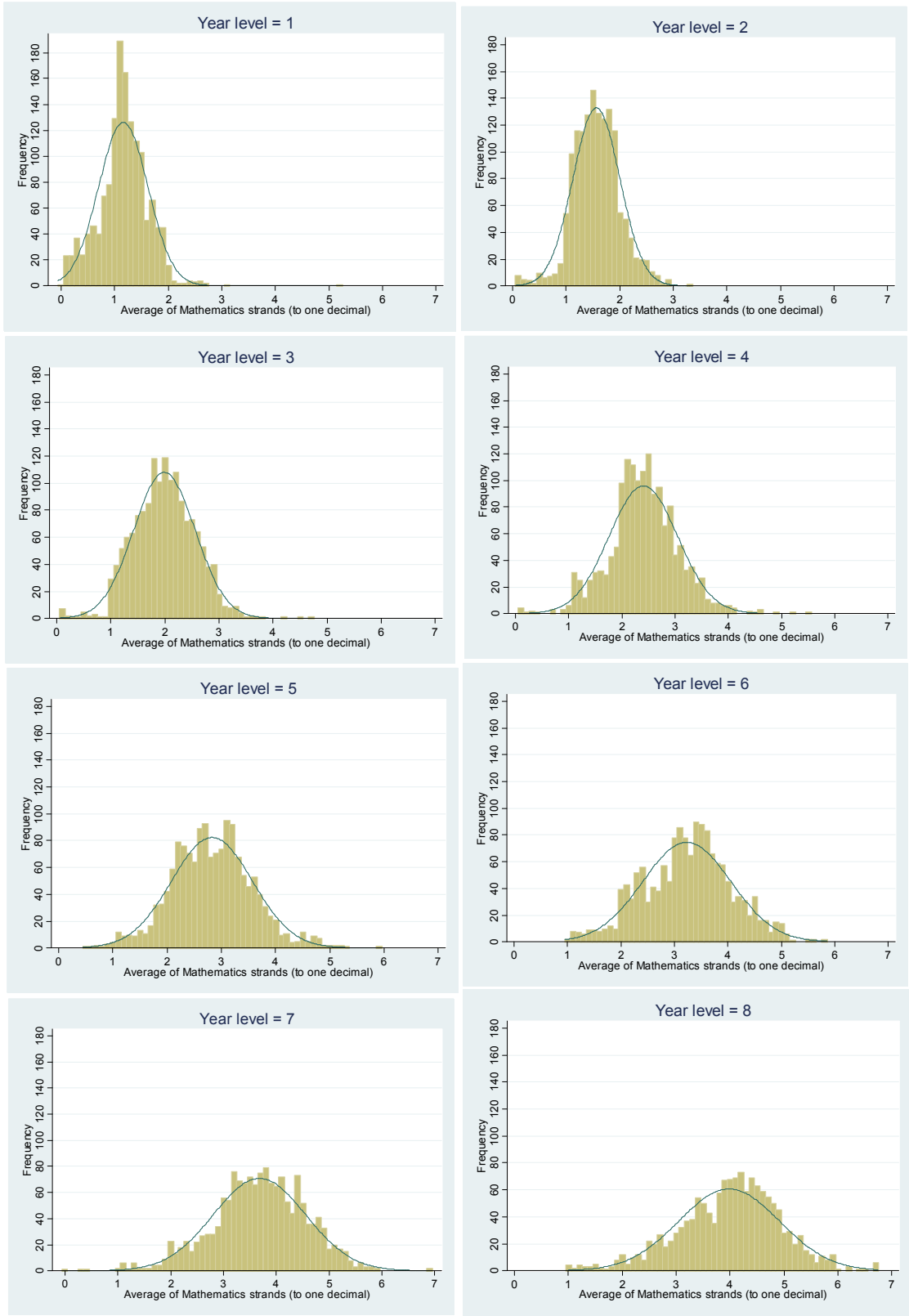


Figure 7.8 Teacher Judgement assessments- Mathematics Learning Area 1998 by strand and Year level

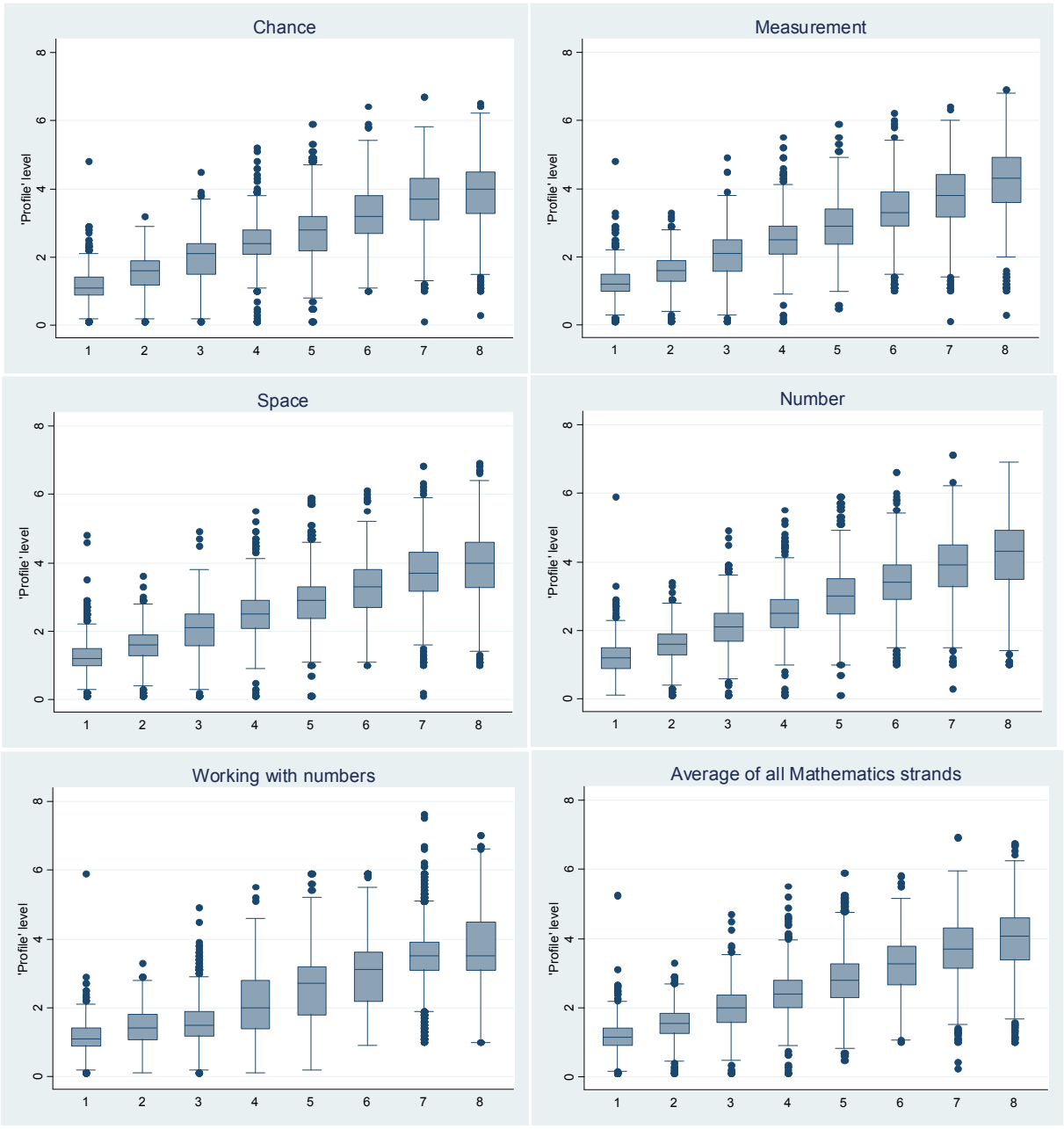
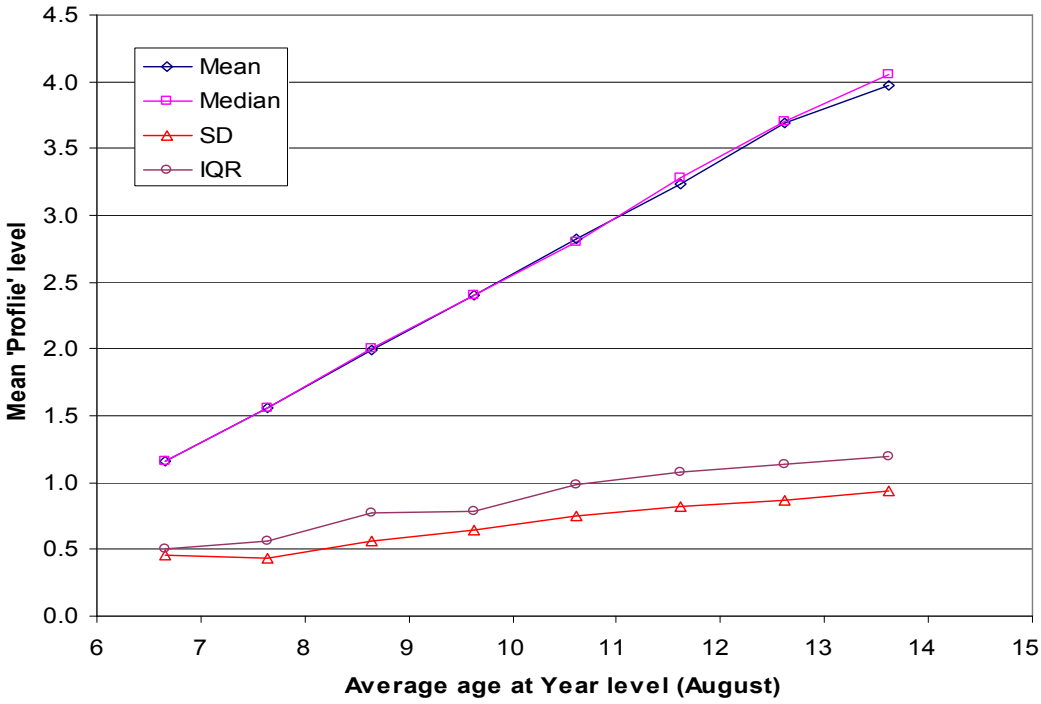


Table 7.2 Mathematics Learning data by Year level –1998: General Statistics

Year level	Average age at assessment (August 98)	Average of all Mathematics Strand values (in Profile units) Mean	Median	SD	IQR	SE of mean	Skew-ness	Kurtosis	N
1	6.66	1.16	1.16	0.46	0.51	0.01	0.45	7.52	1452
2	7.64	1.56	1.56	0.44	0.56	0.01	0.00	3.96	1456
3	8.64	1.99	2.00	0.57	0.78	0.01	0.08	3.71	1537
4	9.63	2.40	2.40	0.64	0.78	0.02	0.09	4.20	1548
5	10.62	2.83	2.80	0.75	0.98	0.02	0.18	3.43	1541
6	11.62	3.23	3.28	0.82	1.08	0.02	-0.08	2.89	1540
7	12.62	3.69	3.70	0.87	1.14	0.02	-0.26	3.34	1554
8	13.62	3.98	4.06	0.94	1.20	0.03	-0.35	3.39	1422
All	**10.14**	**2.61**	**2.48**	**1.17**	**1.74**	**0.01**	**0.36**	**2.50**	**12050**

The consistency of the gradient of improvement in assessed learning indicates that many teachers, over multiple Year levels perceive the mean performance of students as having increased by a constant amount for each year of schooling. The spread of the assessments increases with Year level/age, reflected in SDs and IQRs. As for English, the Year 8 teachers assess students to be, on average, at a lower point than the continuation of the primary teacher gradient would expect.

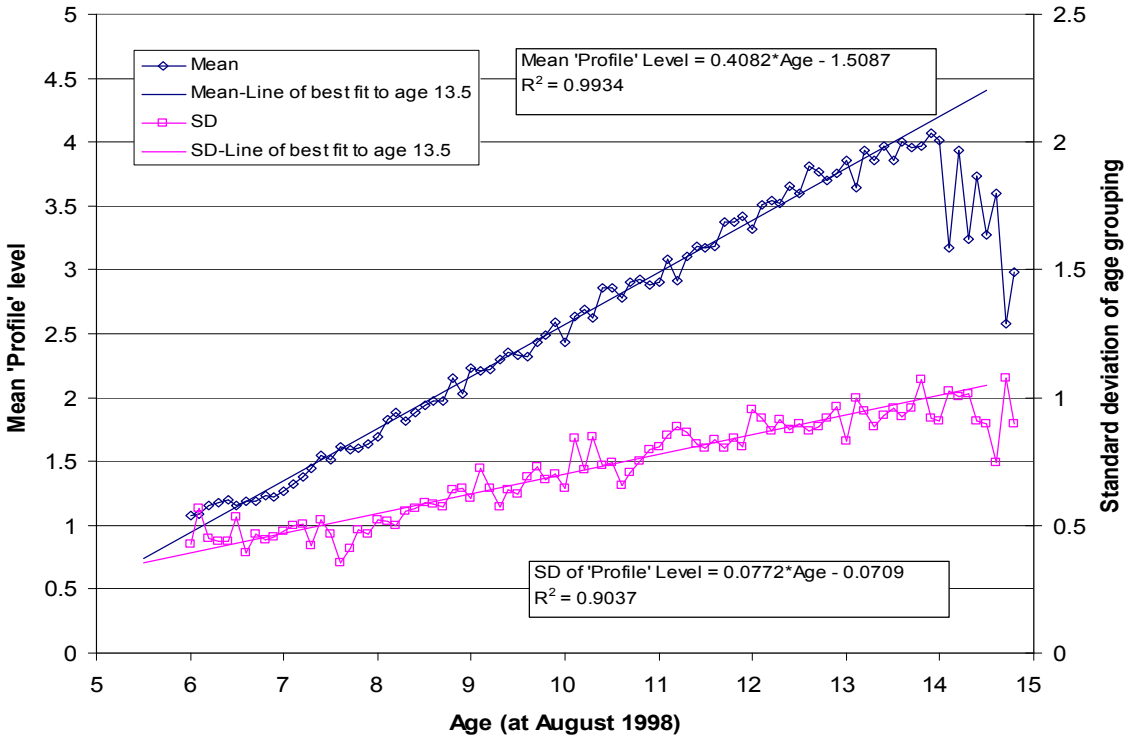
Figure 7.9 Teacher Judgement assessments - Mathematics Learning Area 1998: means, medians, standard deviations and inter-quartile ranges by Year level



Clearly there are different perceptions of learning status by secondary teachers relative to primary teachers. Whether it is a reflection of the actual learning status, a cultural difference between how secondary and primary teachers see learning, or any of a range of other factors

cannot be determined from this data. What is known is that from a test perspective, accepting the general model in Chapter 6, the amount of increase in mean learning status diminishes with Year level and age. This matter is taken up again in Chapter 8.

Figure 7.10 Teacher Judgement assessments - Mathematics Learning Area 1998 Mean profile level all strands combined, by age



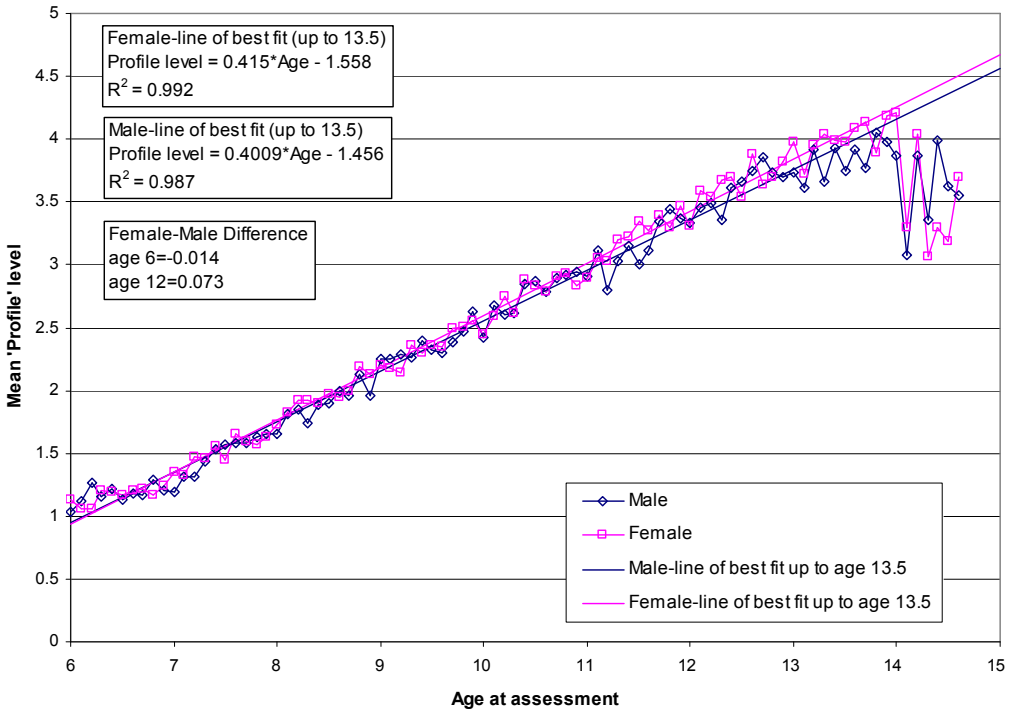
The pattern of higher mean learning status for age groupings at 0.1 of an age appears to hold for the mathematics assessments as it does for English. A straight line regression of mean learning status in profile level units up to age 13.5 has a gradient of 0.4 profile units per year of age, and an R^2 of 0.9934. There is clearly a strong relationship of improving learning status with age. The SDs also show a linear trend with age (to 13.5), with a linear regression as good a fit to the points as a quadratic - in mild contrast to the English data in Figure 7.4 where a flattening of the SD curve occurs after age 12.

Gender differences in the Mathematics Learning Area

Identifying the data by gender (by age), as illustrated in Figure 7.11, reveals a clear difference between teachers' perceptions of mathematics learning compared with teachers' perception of English learning. In Figure 7.11 the trajectories for males and females are intertwined. A regression of the means on age suggests a very slight advantage to females, and a slightly greater variability in the assessment of males (slightly lower R^2 for males). Compared with the clear gender difference found for English, in Figure 7.5 and confirmed in the test model in

Chapter 6, the difference in teachers' assessments by gender for mathematics are trivial. There is a hint that teachers at secondary level might perceive a small mean difference in favour of females, in contrast to the test model which suggests a small mean difference in favour of males. In both cases the differences are very small. Teacher judgment assessments in the Victorian VELS/CSF (Chapter 4) show the same tendency for teachers in higher Year levels to judge female students on average to be slightly ahead of males of the same age in their learning. In the next chapter the test data do not support this difference.

Figure 7.11 Teacher Judgement assessments- Mathematics Learning Area 1998 – Mean profile level of all strands combined, by gender of students



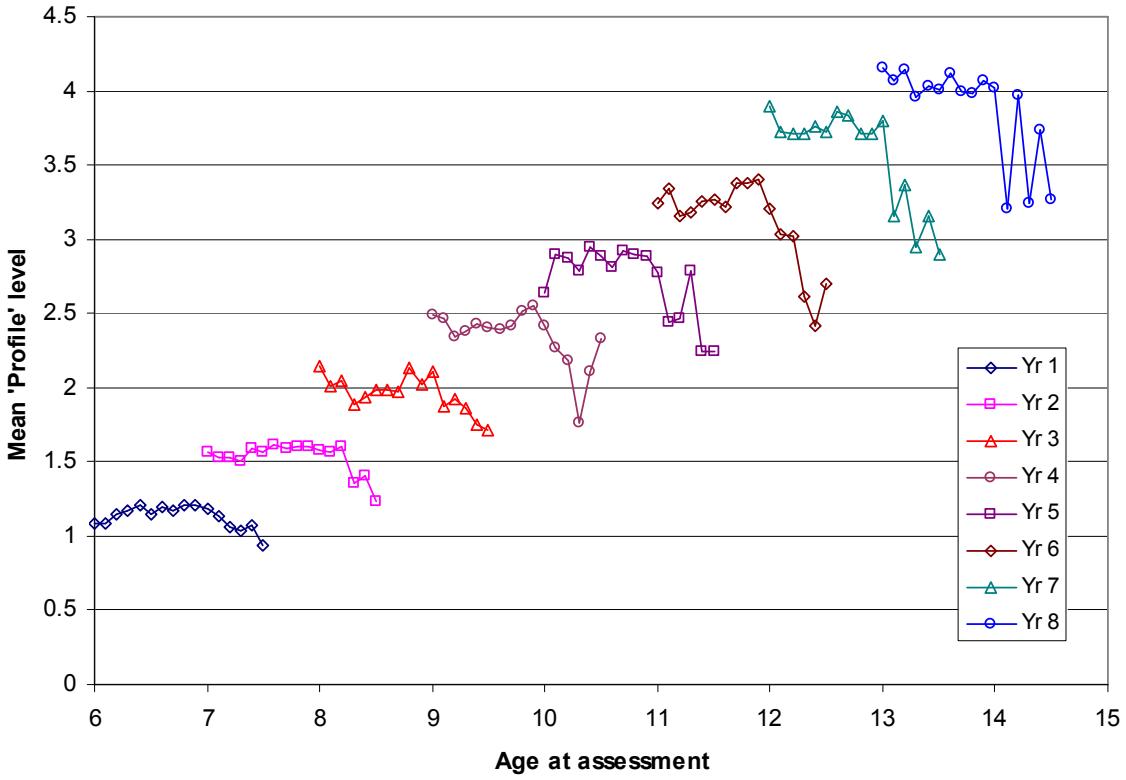
As for English the assessments for students up to age 14.6 are shown for completeness,

The general upward trend applies up to age 14.0. The older students in Year 8 up to age 14.6 are included but a very small number of students older than this are censored. Up to age 14.0 the students are of normal age for the Year level. As shown below in Figure 7.12, older students outside the normal age range for the Year level generate a tail effect. The proportion of students in the tail is small. Data in consecutive Year levels, summarised without considering Year level, do not show the tail due to the cancelling effect of the small numbers of under normal age high scores on the over normal age lower scores, leading to the age means following the general trend.

Within Year level trends by age

The age patterns within each Year level are shown in Figure 7.12. At Year 1 there is a positive gradient with age within the Year level up to the limit of the normal age group for the Year level (7.0 years in August), and then a tail of reducing means for the older students. However the negative gradient of the tail is much less than for the higher Year levels. This pattern is plausible, as mathematic development may not show such a large between-student variation in the early stages. The overall mean for Year 1 (1.16 profile level units from Table 7.2) is very low on the profile level scale and as shown in Figures 7.8 and 7.9, there appears to be a reticence to assess students as "not yet at level 1". These two aspects might explain the slighter tail. In the model in Chapter 6 the tail is similar to that of the upper Year levels, an artefact of being a direct derivation from the Year 3 test shape and the stretching of the distribution on the learning axis to match the steepness of the assumed learning trajectory.

Figure 7.12 Teacher Judgement assessments - Mathematics Learning Area 1998: Mean profile level of all strands combined by age within Year level



The pattern with age over the Year levels (above Year 1) while variable, indicates a general positive trajectory with age within a Year level up to the oldest within normal age point and then a tail for the above normal students. As for English, the data are censored with cases of very low frequency below normal age and above 0.5 above normal age not shown. At Year 8 it would appear that there is no relationship of mean learning status with age, within the normal age range for the Year level; that is the age effect appears to have disappeared. The

older student tail still occurs, reflecting the generally lower mathematics skill of the over-age students.

Common findings across two data collection periods and learning areas

There are some strong common features of the view of student learning development with Year level and age as seen by teachers. The perceived shape of learning for the mean of Year level cohorts is the same whether the learning area is English or Mathematics. The shape holds when the assessments are summarised by age at a highly refined unit of age (0.1) approximating a month of age. The shape is linear with Year level/age until the end of primary school. This is in contrast to the shape indicated by test based IRT measures of ostensibly the same learning, which is curvilinear (Chapter 6).

A linear trajectory is typical in a grade-equivalent approach (Schulz & Nicewander, 1997) hinting that teachers may be using their expectations for a given Year level as the basis for their judgement. Teacher assessments referenced to the Profiles scale, replicated over time and subject, produce similar linear patterns. SDs increase with Year level/age. In contrast, IRT descriptions of learning development with age (Chapter 5) and the IRT related model for tests in South Australia (Chapter 6), the SDs generally reduce with age. The NAPLAN (2009) data for 2008 have this property of diminishing SD for all subjects except Writing.

The implication in the apparent grade-equivalent pattern is that teachers' assessment processes appear to draw on the general perception that teachers have of the standard for the Year level. This applies even though the assessments are strongly referenced to the level criteria framework to assist in the judgement of where in the framework a given student is placed. It is remarkable that this grade-equivalent standard, averaged over about 180 teachers per Year level in 1997 and 300 teachers per Year level in 1998, has such a strong linear relationship with Year level over the full primary Year level spectrum. The trend holds across the two subjects under investigation with slightly different gradients and based on general patterns found for the 6 other learning areas surveyed (Rothman, 1998; DECS Curriculum Bulletins 1998-1999), the same trend appears to hold generally across the other six learning areas.

Teacher judgement assessments of student learning have two subtle elements. The first is the apparently strong relationship of learning with decimal age. It is assumed that age differences of the order of a month of age, were not considered by teachers in making their assessments. The age effect applies to the data undifferentiated by Year level as well as within Year levels, at least below Year 7.

The second element is the clear perception by teachers that females, on average, at a given age have a slight performance advantage over males in English language development. Contrasting with this is the different perception of mathematics learning where there appears to be hardly any gender effect (notwithstanding the consistent but subtle difference at higher Year levels between test and teacher judgement assessments). It is unlikely that gender would have been consciously considered when the assessments were being made. The gender differences for English, when compared in Chapter 8 with the patterns identified by test summaries, follow approximately the same patterns implying the effects are not merely the result of a bias. Slightly higher results for females, however, appear to apply in teacher judgements of Mathematics at higher Year levels.

The data also indicate that teachers generally view learning at a fine degree of refinement. How well this refinement matches the scale from the testing process generally and for individual students is addressed in the next chapter. It is observed here that teachers used the full spectrum of available points (at 0.1 of a Profile level) for the population of students in a Year level. The response format (clicking on a continuous line) did not allow a teacher to see that any judgement was at this level of refinement nor is it possible to resolve from this data collection how teachers would feel about applying a numerical value to learning progress. However the principle is clear that teachers might be able to estimate learning status at the refinement of about the equivalent 2 to 4 weeks of learning development (though not necessarily at this implied frequency).

Acceptability of teacher judgement assessment to teachers

As documented in Chapter 3 the introduction of the Statements and Profiles for Australian Schools (SPFAS), and its precursor in South Australia, achievement levels, was controversial and led to teacher industrial concerns about workload and the possible misuse of assessments. Teachers' confidence in the their judgements was considered to be important to establish and thus data on confidence was collected as a condition for union and teacher participation. Teachers were asked two confidence related questions at the completion of their judgements: teachers' confidence in the process generally and their confidence in the assessment for each student individually. A five-point rating scale was used with a rating of 5 indicating the greatest confidence.

Table 7.3 reports the percentage of responses at each point on the 5-point scale for both questions, along with the 'no response' rates. Confidence in the general process appears to be less than the confidence in the assessment for individual students, based on the combined percentage of responses at rating points 3, 4 and 5. Confidence in the process was around 60%, assuming that the highest ratings of 3 to 5 reflect a positive view of the process. The

more specific confidence in each individual student assessment was higher. Over the combined ratings 3 to 5, confidence was around 70%. Given the industrial concerns about the introduction of the assessment reporting and the directly-connected national curriculum (covered in Chapter 3), the confidence ratings appear surprisingly positive.

Table 7.3 Ratings by teachers of their confidence in the process and in their specific assessments.

	Confidence in the process		Confidence in specific student assessments	
	1997	**1998**	**1997**	**1998**
No rating	20.7%	32.3%	20.6%	18.6%
1	7.3%	2.9%	5.9%	3.5%
2	7.9%	6.0%	5.7%	7.5%
3	17.5%	22.8%	17.9%	25.2%
4	29.5%	27.2%	31.3%	34.5%
5	17.1%	8.7%	18.6%	10.7%
3+4+5	**64.1%**	**58.7%**	**67.8%**	**70.4%**

Concluding comments

The introduction of a profiles approach to curriculum and assessment in schools in South Australia was driven by curriculum leaders rather than assessment advocates. The focus of curriculum leaders was essentially and appropriately on how the general profiles framework might help schools and teachers match and refine their existing curriculum arrangements to the described structure of developmental learning. The relationship of the statements and profile structures to assessment and recording of student learning was not fully addressed. The ways in which individual and personal assessments of learning status using a developmental map might be applied to add refinement to a profile level assessment were not often considered. The purpose of the collection of data was not clearly resolved and delayed a number of years. The mechanics for the collection were addressed only in the negotiations to conduct the collection (as described in Chapter 3). In particular the level of resolution possible for a teacher judgement assessment was not considered in a way that could have led to pretested process for the possible degree of resolution. As a result the collection of data was an imperfect process.

A key concern was the perception by teachers that a profiles-based assessment might apply only irregularly. For many teachers, assessments were made only for the two data collections and not embedded as general classroom records. This is unsurprising given that the record of learning status was of such low resolution that keeping records would be seen as a waste of teacher effort. A personal student history in profile level units would not contribute to the day-to-day learning support for students. However, the data summarised above suggest that teachers' judgements provide a very comprehensive overview of what learning development

looks like over 8 years of schooling. The consistency of the annual increment in the means of teacher judgement assessments suggests that some powerful underlying perception of learning growth is understood by a sufficient number of teachers to ensure the means grow as observed. Not all teachers need to have this ability for the pattern to exist, and persist, over two collection periods. However it seems for all 8 of the learning areas described in the curriculum description (SPFAS) the same general linear trajectory pattern of the mean of the Year level assessments applied.

The data collected in 1997 and 1998 indicate that teachers can articulate an on-balance judgement assessment using a well-described logical framework, even if the framework has many ambiguities. The allocation of a numerical value to learning status, notionally the same process that applies with formal testing, is feasible and, in operation, the data generated are consistent with what is expected in learning development. The major difference, in broad terms, between learning development as seen by teachers and as described by vertically scaled IRT tests is the shape of the general trajectory of learning. Teachers using the SPFAS framework generate data that describe a linear trajectory with increasing spread. IRT data describe trajectories that are non-linear, with mean growth per period reducing as upper segments of the learning scale are approached. The spread of students around the trajectory reduces rather than increases. What is not clear from examining the teacher data in isolation is the extent to which assessments of individual students by teachers and tests produce equivalent assessments for a student. The next chapter explores a range of ways that data above and data from test sources can be compared as grouped data and for individual students.

Chapter 8 Teacher and test assessment compared

> … such is the hegemony of traditional psychometrics, that these alternative assessment systems are widely characterised as 'soft' and 'unreliable'. The pioneering work of our best teachers has run far ahead of the available theory, and I believe the lack of theoretical support from the academic community for these innovative practices has made it much easier for politicians to deride and dismiss any assessment practice that does not meet their own aims.
>
> Wiliam, 1994, p. 17-18.

This chapter applies a range of methods to convert teacher judgement assessments to the scale of test assessments so that teacher / test comparisons can be made. It is already clear from the different shapes of the trajectories of the IRT based test view (Chapter 6) and the profiles based teacher view (Chapter 7) that the conversion of one assessment process to the scale of the other cannot be a simple linear transformation. The purpose in comparing the two assessment processes is to establish the degree to which they were interchangeable in 1997 and 1998. The chapter considers a variety of methods for equating the scales of the two processes. Once the scales are approximately equated it should be possible to understand the degree to which judgements made by teachers can be considered as equivalent to the scores provided by tests. The validity of the assumption that teacher and test assessments can be compared is also considered.

Equating Teacher and Test scales

Assumptions

Prior to addressing the process for bringing the two assessment arrangements to a common scale, some discussion is required of the validity of the assumption that the two processes are addressing the same dimensions in each case (English compared with Literacy, Mathematics compared with Numeracy). The equating of the teacher judgement and test scales within specific learning areas is logical only if the case can be established that both processes are assessing the same dimensions of learning. The broadness of the English and mathematics traits and the commonsense understanding of these, imply a strong likelihood that the order, at least, of students on each scale could be expected to be similar whether teacher or test assessment is used. As part justification for continuing the equating processes, the correlations of the scores on the two scales are 0.659 (n=1275) for English/Literacy and 0.57 (n=2105) for Mathematics/Numeracy (see Tables 8.1 and 8.2 later). These are lower correlations than are expected in parallel test forms, but indicate a reasonable degree of consistency of order from the two assessment processes.

The assumptions that the English learning area and Literacy tap the same underlying latent trait of learning or that the Mathematics Learning area and Numeracy are similarly derived, are not unique to this study. As described in Chapter 4, the Victorian student assessment system has operated on this assumption in very similar circumstances. However a much closer alignment of the test to the common curriculum framework applies there as both the test and teacher judgement assessments relate directly to the VELS/CSF frameworks. The teacher judgement assessment frameworks were very similar in SA and Victoria in 1997 and 1998; and, based on the author's observation, the test items, while developed to different specifications, appear to be similar in style.

Based on the precedent actually operating in Victoria, there is justification for assuming that the underlying latent dimensions of the tests and those for teacher judgement assessments are similar enough to explore converting data from both sources to a common scale.

Data sets used

The initial step in the equating of the teacher profile level scales with the test scales is the matching of records from the test files with records from the teacher assessment files to find students in common. Using the student identifier used in Chapters 6 and 7 to assign dates of birth, the author compared the records in the teacher-assessed file with the test files (student identifiers now added) to find cases common to both files. As described in Chapter 6 only 72% of test cases were assigned a date of birth for 1997 and 75% to 80% for 1998, depending on the test. Of the sample of students assigned teacher judgement assessments, 64% of cases in the combined Year 3 and Year 5 set for 1997 were matched (n=1275), and 68% for 1998 (n=2105) (Table 8.3). The proportions of the test cohorts who were also assessed by teachers using the SPFAS approach were 5% in 1997 and 9% in 1998. The matched cases may not be randomly selected from their parent distributions.

The teacher assessments

The strand scales used by teachers have eight levels of development from the beginning of school to Year 12. Ten scale positions within each of the eight levels, obtained from teachers clicking on a progress line, were used as progress indicators within a level. As a result each strand has a range of 90 score points from 0 to 8.9. Scale positions used for Years 1 to 8 only are in the range 0 to almost 7.0, that is approximately 70 scale positions are used across these 8 Year levels. For the Year levels 3 and 5 only, most assessments are in the range 1.0 to 5.0, although the full range is 0.2 to 5.9.

In the English learning area, the correlation of Speaking and Listening with the test at Year 3 is lower than for the other two strands (Table 8.1). However the correlations of Speaking and

Listening with the other two teacher-assessed strands are high (0.92 and 0.93). These between teacher judgement strand correlation values are not reported in Table 8.1.

Speaking and Listening was not included in the averaged scores analysed in Chapter 7 on the grounds that the tests, with which the teacher judgements were to be ultimately compared, covered only Reading and Language, the latter in written forms only. That position is modified in this chapter. All three strands are used in the Rasch analysis to make the analysis feasible. Using three strands allows three items for each student.

The correlations of the strands with the test at each year level separately and as a combined data set are shown in Tables 8.2

Table 8.1 Correlations of English teacher assessments with Literacy test assessments – 1997

N=1275	Teacher assessed- Reading	Writing	Speaking & Listening	Correlation with the average of all three strands
Test Year 3-Literacy	0.55	0.52	0.39	**0.52**
Test Year 5-Literacy	0.60	0.59	0.51	**0.60**
Test Combined Years	0.67	0.65	0.58	**0.66**
Female –combined	0.67	0.66	0.58	**0.66**
Male -combined	0.65	0.62	0.56	**0.64**

In the mathematics learning area the Working Mathematically strand correlates between 0.75 and 0.78 with the other four mathematics strands , while they correlate with each other in the range 0.85 to 0.9. It is assumed that Working Mathematically is either measuring a different dimension to some extent or was less well understood by teachers. Either way Working Mathematically appears to be different to the other strands. As a result the Working Mathematically strand is not included in the averaged score for each student in the analysis in this Chapter, nor as an item in the Rasch analysis described later. The test-teacher correlations by strand are shown in Table 8.2.

Table 8.2 Correlations of Mathematics teacher assessments with Numeracy test assessments-1998

N=2105	Teacher assessed- Chance	Measure-ment	Number	Space	Correlation with the average of all four strands	(Working Mathematic-ally)
Test Year 3-Numeracy	0.37	0.40	0.42	0.39	**0.42**	(0.32)
Test Year 5-Numeracy	0.48	0.46	0.49	0.47	**0.50**	(0.37)
Test Combined Years	0.53	0.55	0.57	0.54	**0.57**	(0.47)
Female –combined	0.51	0.52	0.54	0.52	**0.55**	
Male -combined	0.56	0.57	0.59	0.56	**0.59**	

The correlation coefficient for the combined data set of the test and teacher assessments for English/Literacy is 0.659. For Mathematics/Numeracy the correlation of the combined data set teacher with test assessments is 0.571. These values are less than usually expected for parallel forms of test-based assessments.

Comparing raw scores

A range of statistical characteristics of the Year 3 and Year 5 cases with both a teacher and test assessment are reported in Table 8.3.

Table 8.3 General Statistical Characteristics of common cases of Teacher assessments and Test assessments, 1997 and 1998

		1997 Teacher (English) Profile units	Test (Literacy) Logits	1998 Teacher (Mathematics) Profile units	Test (Numeracy) Logits
Year 3	**Mean**	2.20	0.50*	2.13**	0.17
	Median	2.20	0.61	2.13	0.22
	SD	0.55	1.26	0.55	1.41
	SE (Mean)	0.02	0.05	0.02	0.04
	Min	0.17	-6.42	0.50	-6.18
	Max	4.00	3.80	4.65	4.58
	Skewness	0.09	-0.24	0.06	-0.24
	Kurtosis	3.39	3.97	3.14	5.46
	N	702	702	1035	1035
Year 5	**Mean**	2.98	1.74**	2.96**	1.35
	Median	3.00	1.78	2.95	1.38
	SD	0.67	1.19	0.71	1.22
	SE (Mean)	0.03	0.05	0.02	0.04
	Min	1.13	-2.06	0.60	-5.40
	Max	4.97	5.36	5.90	5.96
	Skewness	0.14	-0.21	0.25	-0.68
	Kurtosis	3.44	3.44	3.61	8.18
	N	573	573	1070	1070
Combined	**Mean**	2.55	1.06	2.55	0.77
	Median	2.50	1.07	2.50	0.83
	SD	0.72	1.37	0.76	1.44
	SE (Mean)	0.02	0.04	0.02	0.03
	Min	0.17	-6.42	0.50	-6.18
	Max	4.97	5.36	5.90	5.96
	Skewness	0.33	-0.21	0.40	-0.46
	Kurtosis	3.24	3.45	3.37	5.27
	N	1275	1275	2105	2105
Year 3 assessments (matched and total)		702/1005 (70% matched)	702/12437 (6% matched)	1035/1537 (67% matched)	1035/12794 (8% matched)
Year 5 assessments (matched and total)		573/996 (58% matched)	573/11973 (5% matched)	1070/1541 (69% matched)	1070/12471 (9% matched)
All cases (Year 3 + Year 5)		1275/2001 (64% matched)	1275/24410 (5% matched)	2105/3078 (68% matched)	2105/25265 (9% matched)

* Difference from means in Tables 6.5, 6.9, 7.1 or 7.2 significant at 5% level based on t-test

** Difference from means in Tables 6.5, 6.9, 7.1 or 7.2 significant at 1% level based on t-test

The table indicates the general statistical characteristics of the students common to both assessment processes. Listed are teacher assessed profile level values (based on the averaging

of strands for each student) and test values as logit scores obtained from the item linked vertical scale for the tests. The general statistics for the full test populations are found in Tables 6.5 (1997 test) and 6.9 (1998 test) and in Tables 7.1 and 7.2 for the teacher assessments.

The means and medians in each column of Table 8.3 are close to those in each of the 8 separate Year level samples and the 4 combined samples for all but the 1997 Year 3 Literacy test where they differ by 0.09 of a logit. The skewness and kurtosis statistics are generally comparable to the original data sets in Tables 6.5, 6.9, 7.1 and 7.2. The test means for 1998 are not significantly different from the full cohort but the means for 1997 differ by up to 0.2 logits, and are different beyond the 5% and/or 1% significance levels. The means of the sub-samples with both teacher assessments and test assessment compared to the original full sample teacher means for 1997 are not significantly different. For 1998 the sub-samples with teacher-and test assessments have means approximately 0.1 of a profile level above the full sample means. The SDs in the sub-samples of teacher assessments with matched test cases are comparable to those for the full test cohort in Tables 6.5, 6.9, 7.1 and 7.2.

That the students with both test and teacher assessments do not have means identical to their parent samples i.e. the selection may not be random, is not regarded as critical. The purpose in this first stage is to examine the assessment scores only for the set with both assessments, in an attempt to equate the teacher judgement assessment scale to the test scale. Necessary for such a process is a good spread of cases on both the test scale and the teacher scale. The common cases provide this spread.

The assessments are from multiple teachers. The set of records where a test and teacher assessment can be compared depend in the first instance on the allocation of the student identification codes to each file. For various reasons not all students who were assessed by both processes could be identified. As a result some teachers who provided assessments may have been removed, or at least part removed, from this part of the analysis. While it is impossible to know which teachers were affected, an unintentional bias in the deletions might have occurred. However, for the purpose of the analysis the failure to allocate student identifiers is assumed to be random, or, at least, trivial. Two further issues arise that are important in the appropriateness of equating the scales. These are addressed in detail in Appendix 11. For 1997 the data for teachers were obtained up to three months later than the test assessments. It has been assumed for equating purposes that this time difference does not exist. As the process is to establish a relationship for a one-off analysis, this time difference is immaterial to the results. In reality the conversion relationship should be set such that each profile value takes a slightly lower test scale value (of the order of 0.1 logits lower). For the 1998 data, tests and teacher judgement assessments were at almost the same time. The second

issue is the difference in the criterion used by the test (50:50 odds) and that likely to be used in teacher judgement assessments where mastery of a skill is required. For the teacher mastery of a behaviour or skill might require expression at 80% to 90% of the time. This important difference is not adjusted for in the analysis and is discussed in more detail in Appendix 11.

The nature of the assessments as continuous or discrete data also needs comment. Teacher judgement assessments can be considered as approximately continuous through the averaging process applied across strands. Test data are scaled on a Rasch scale at points expressed to two decimal places. Even though the actual data points bear a direct relationship to the original number correct scores (i.e., they are discrete), they are assumed to be continuous for the purpose of the explorations. The concentration of the points on the test scale as shown in scatter diagrams, highlight that the test data points are not continuous in practice (Figures 8.8 and 8.9).

Equating approaches

The principles of the range of equating processes used are summarised in Appendix 11. The processes described include mean, linear and equi-percentile equating. All are used in sections of the analysis. A specific linear equating approach (non-anchored Rasch scaled linear equating) is used as one basis to equate the teacher and test scales. In this process the two scales are developed independently for all the teacher and test cases using the Rasch model. Then for the common students the means and SDs are equated. Based on the summaries by Year level for the teacher-assessed cases presented in Chapter 7, the relationship of mean learning status with Year level appears to be linear. From a test perspective the trajectories of mean learning status are curved, with growth increments diminishing with Year level (Chapter 6). As a result, the arrangement to convert teacher assessment scores into the framework of the test over the full range cannot be linear.

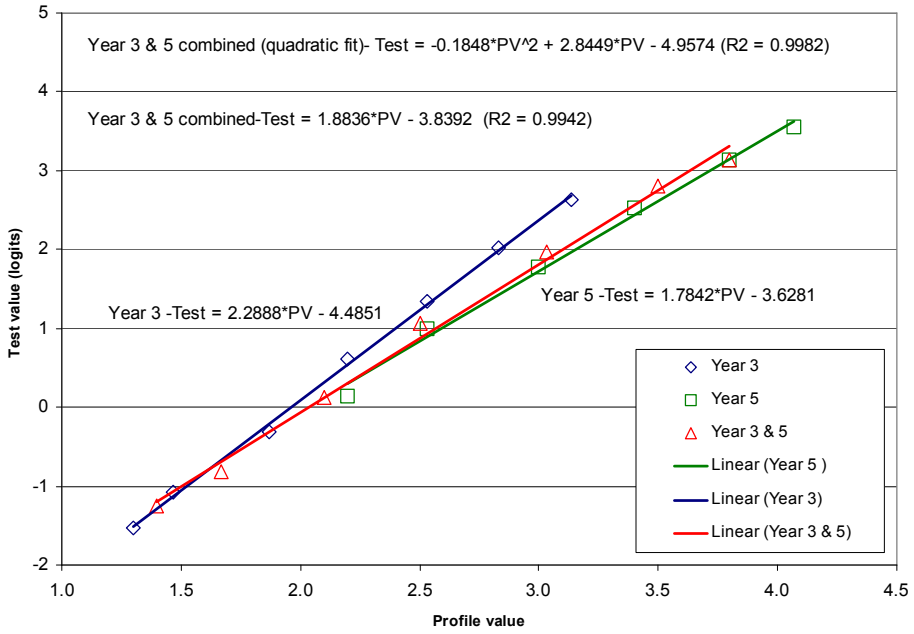
The next section of the chapter establishes the similarity of the equi-percentile equating result applied to the combined Years 3 and 5 to a Rasch model supported equating process. This comparison is made to justify the subsequent use of the Rasch model even though its application to the limited teacher-judgement data is problematic. That both equating processes produce similar results for the most part along the teacher judgement scale is offered as evidence that the conversion process of teacher judgement assessment scores to test scale values is robust.

Equi-percentile equating: Year 3 and 5 cases separately and combined.

Equi-percentile equating of the two scales, using the students common to both forms of assessment, is the simplest equating process to apply. Figure 8.1 illustrates the results of

equi-percentile equating for Years 3 and 5 separately. Using the seven percentile points (5, 10, 25, 50, 75, 90, 95), the Year 3 test scores are plotted against the mean profile level per student. These points turn out to have a clearly linear relationship. Ordinary least squares (OLS) regression is applied since R^2 is virtually 1.0, meaning that regressing Test on Teacher assessments produces the same result as regressing Teacher on Test. The Year 3 fitted line has a gradient of 2.28. Similarly the Year 5 percentile points are approximately linear with a gradient of 1.78.

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.



The graphs suggest one of the potential contributors as to why the general relationship of the test scale to the profile scale is curved. Within a Year level the equating relationship is linear but the gradient of the relationship is diminishing with increasing Year level. It is a leap from the data of the two known Year levels to assume the likely gradients at other Year levels. At Year 2, however, it might be assumed to be steeper. For the intermediate Year 4 a gradient between those of Years 3 and 5 is logical. The same apparent variation of gradient with Year level applies independently in the mathematics data collected one year later.

It is known that as Year level increases, the span of the development range of students increases (Chapter 7) based on a teacher judgement assessment scale view. The SDs increase with Year level. The bulk of the class could be expected to be placed around the Year level mean, the judgement of which aggregated over many teachers is linearly increasing with Year level, as is the spread (see Figures 7.3, 7.9). As the spread increases, the learning-status-estimates further from the Year level mean are likely to be made with less detailed knowledge of the typical skill level of students at the extremes by teachers at that Year level. From

Figure 8.1 a Year 3 teacher sees a student with a test score of 2 logits (above the 90th percentile for Year 3) as at about 2.8 profile level units. The same test score position, now only above the 75th percentile for the Year 5 teacher, is seen as 3.2. The teacher assessment scales are Year level specific.

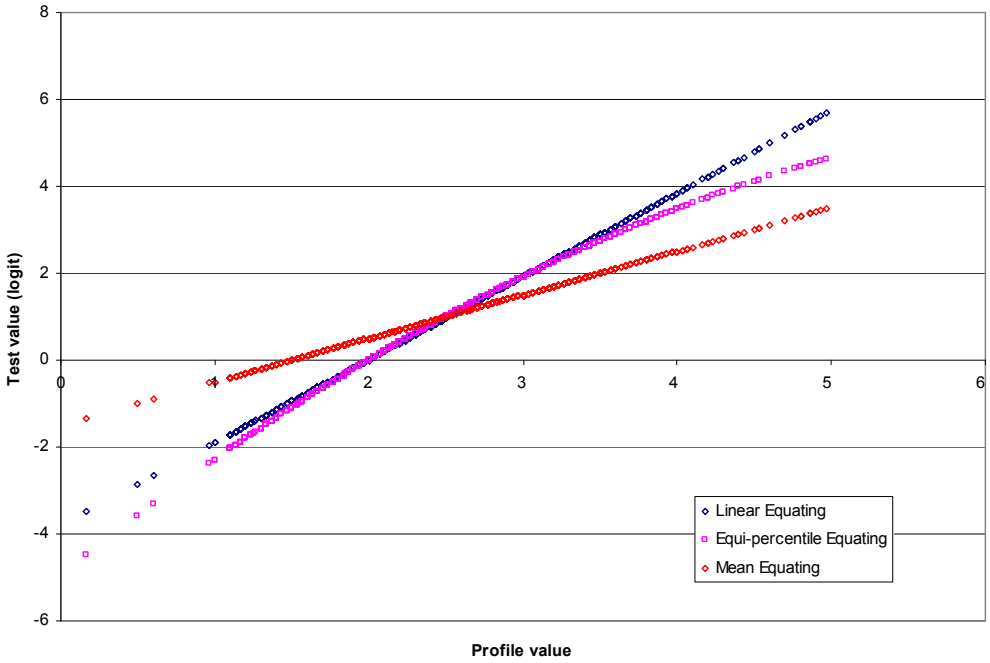
The purpose at this point is to estimate the general profile level to test score relationship over the profile level range from 1 up to 6 using the equi-percentile approach, even though the data for common students covers the range of 0 to 4 profile levels only. The common cases at Year 3 and 5 are well balanced on the spectrum of Year levels from 1 to 8 and thus profile levels from 0 to 6. The combined Year 3 and 5 teacher data sets (Table 8.3) have improved correlations with the test scores, relative to the two Year levels treated separately. The correlations are illustrated in Tables 8.2 and 8.3. For simplicity, the analysis continues on the basis of using the combined data for Years 3 and 5 to estimate the general equating relationship of profile level scores to test scores. An acknowledged consequence is a mild distortion of the equating relationship at each Year level. It will be shown later that the Rasch model equating produces a relationship for the two scales similar to that of the combined Year level equi-percentile method. As a result the choice of preferred process depends on the other benefits that might arise from the equating process chosen.

Using the combined set, based on equi-percentile points, the marginally best fitting equating relationship is curved. Figure 8.1 provides the linear and quadratic expressions for the lines of best fit. The curve itself is not shown to reduce visual complexity. The R^2s for the linear and quadratic fitted curves are virtually identical but the quadratic fit has a slightly higher R^2 value. Since the relationship of profile units to test score units is already known to be non-linear, the curvilinear relationship should be preferred for extension outside the Year 3 to 5 range. Ultimately a Rasch model equating is adopted as described below. It will be shown to be approximately identical to the equi-percentile equating but, prior to arguing the advantages of that process, three equating processes are considered to illustrate the ways in which those results differ. When it is shown empirically later that the non-linear equi-percentile approach approximates the Rasch model, the Rasch model equating result can be seen as proxied by the equi-percentile solution.

Comparing the equating results from mean, equi-percentile (linear) and equi-percentile (non linear) relationships.

Figure 8.2 illustrates the application of three equating processes to the 1997 data for Years 3 and 5 combined, based on plotting the result for each of the 1275 data points.

Figure 8.2 1997 Profile to Test scale equating by equating method, Year 3 and 5 data combined-English



As would be expected the mean equating process coincides with the other approaches only at the means of the two scales. The lack of equating of the spread generates an inadequate solution. The equi-percentile (linear) equating and equi-percentile (non-linear) equating produce approximately similar solutions over the range of 1.5 to 3.5 profile units. Outside this range the equating relationships spread apart. Of the three processes, only the non-linear equi-percentile equating is sensitive to the non-linear relationship between the teacher and test scales identified in earlier chapters.

Rasch model equating

An alternative equating process is applied, based on a Rasch model analysis of the full teacher assessed cases from Years 1 to 8 for 1997, 7871 cases altogether. In this process the three assessment strands in English are regarded as items. The item score values are obtained by deleting the decimal point. The items can take a value from 0 to 89, although the highest actual score is 70. The analysis is at the low limit of tolerance for a Rasch model using Winsteps and is not a conventional analysis. The approach uses the Rasch model akin to Wright's (2000) application of Winsteps to multiple regression (Bond & Fox, 2007, p. 203) and takes advantage of the capability of Winsteps to take 99 values for an item when a two-column format is used.

Three general options are available for the equating of the teacher assessments to the test scale under this process. One option is to use the 1275 common students (for the 1997 data) as person anchors for the full 7871 cases. The second option is to use a subset of the 1275

common points as anchors. The third option is to analyse the data set without anchors and then equate the teacher scale to the test scale using linear equating, effectively rescaling the teacher assessment logits to match the test logits.

Testing of all three options indicates that they produce approximately the same general equating result for the non-anchored points. However the first two options fix the student score on the teacher assessment to the test score (a logical expectation of the concept of anchor) for all or some of the points. In both options the anchoring pre-determines that the anchored cases will maintain their original relationships. This defeats the purpose of the investigation of the extent to which the two assessment process produce similar results since some (or all) cases have a predetermined relationship. For this reason, the equating exploration is developed using the third option, without anchors.

Non-anchored Rasch model equating

Appendix 12 contains the detailed statistical summaries of the fitting of the teacher data to the Rasch model for both the 1997 and 1998 collections. As indicated above the application of the Rasch model to such a messy data set is unconventional and produces a large number of poorly fitting cases. To complete the analysis using the Masters partial-credit model (Linacre (2000, p 300), a relatively large number of iterations were required (741 for 1997, 275 for 1998). In this less conventional approach there are 90 categories of partial credit for three items in the English case and for four items for the mathematics case.

In Appendix 12, Table A12.1 the mean square infit value for the three items for 1997 is 0.94. The actual infit values for the three items are 0.72, 0.91, and 1.18. In general terms these items fit the model (between 0.7 and 1.3) and have an item reliability of 0.92. The limited number of items and the person infit mean-square mean well below 1.0 indicates a high degree of over-fitting cases. Boxplots of the distribution of person infit mean-squares are shown in Appendix 12, Figure A12.1. These illustrate the high degree of skew towards 0 with the median well below 1 and the wide spread of values. As Year level increases the spread increases. Inspection of the cases at each end of the spectrum offers some understanding of the reasons. Those cases with infit values at or near zero have no between-strand variation, a consequence of teachers seeing the progress within each strand as equivalent. Cases at the upper end (higher infit values) have one strand where the value varies by 0.5 profile units or more from the other two. The Year level trend reflects the increasing variability between strand assessments as Year level increases, that is the teachers discriminate more between developmental status in each strand as the Year level of the student increases.

Table A12.3 shows high negative residual correlations between items, further evidence of over-fit and the variance in the data explained by the model measure is very high at 97.7% (Table A12.4) supporting the purported unidimensionality of the data. There is lack of randomness in the data due to the few items and commonly high correlations in the ratings for each person on each item due to their being at roughly the same point of development on each strand. Based on Linacre (1999) "some randomness is needed in the data in order to construct a measurement system…In the case of local independence, however, the fit interpretation is reversed. The closer the data comes to the perfect Guttman pattern the less local independence there is, and so the worse the fit." (Linacre, 1999, p. 710) In this application of the Rasch model the purpose is to approximate fit to the model sufficiently to bring the teacher assessment data into an arrangement that assists the transformation of scores from the teacher scale to the test scale. Given the very few items, can the use of the model be justified?

It will be shown below that the relationship estimated for the cases coincides for large segments of the teacher scale with the equi-percentile (non linear) solution. On this basis the measure values estimated for items and persons are proposed as having sufficient heuristic value to explore the modelled data further. One reason is the benefit of the Rasch model in estimating measurement error for each case.

The comparable set of Rasch model fit statistics for the 1998 (mathematics) data are provided in Table A12.5. More items are provided (4 as against 3) with an infit mean square mean value of 0.75 but a narrower range of infit values. Item reliability is reported as 1.00 with a much higher separation statistic of 24.57 (compared to 3.47 for English). The boxplots in Figure A12.2 show a high degree of overfitting, consistent with the lower infit mean square mean relative to the English data. The SD of the infit is less; the reduced spread is observed in the boxplots, as is the trend of increasing spread with Year level. High negative residual correlations are found (Table A12.7) but with lower values than for English. Table A12.8 indicates support for the premise of unidimensionality but with low local independence as for English, due to the inter-related nature and limited number of items.

Item/Strand difficulties

The strand difficulties for the two teacher judgement collections are shown in Figure A12.3. While the difficulties are presented side by side the scales should not be assumed to be the same. The figure illustrates the closeness of the difficulties of the three items for English and the greater spread of the difficulties of the mathematics strands. The appendix provides more comparisons of strand issues that are not dealt with here. Based on the test-teacher correlations reported earlier (English-Literacy 0.66, Mathematics-Numeracy 0.57- both highly significant at the 1% level), continuing the analysis using the total test score and profile average over strands has reasonable face validity even though the teasing out of strand detail might be problematic.

Converting the Unanchored Teacher Rasch measures to the test scale

The Rasch measures obtained in Appendix 12 are converted to the test scale by the following process. Using the 1997 data, the teacher-assessed cases common to the tested population (at Years 3 and 5) are selected. The mean of the teacher assessed score for this group has a value of –1.64 logits and an SD of 1.33 compared to the test assessed scores that have a mean of 1.06 logits and a SD of 1.37. These are found in Table A12.2. The teacher measures from the Rasch analysis are converted to the test framework by making the mean and SD of the common cases equivalent to the test mean and SD by the standard procedure (teacher means rescaled to the test mean and spread in proportion to the ratio of the SDs to make the means and SDs of both data sets identical). The result is confirmed in the third column of Table A12.2. The full teacher assessed set, that is the additional students at other Year levels, are re-scaled on the same basis to create a set of 7871 cases with a mean of 1.23 logits and a SD of 2.07 logits. This compares with the original values of –1.47 and 2.00 in the second column.

This process has re-scaled the length of the teacher assessment scale logit. Because it is used as key part of the comparison of the individual common cases, the error of measurement is also rescaled to the test logit scale. In the case of the 1997 data this re-scaling of the error makes little difference to its values (see the right column). However the same process applied to the 1998 data (see Table A12.6) produces an increase in the error of measurement (from a mean of 0.18 for the common cases to 0.27 when rescaled) due to the differences in SD (1.44 test, 0.99 teacher).

The results of the rescaling are shown in Figures 8.3 and 8.4. The original teacher Rasch measures in logits on the vertical scale are summarised at their profile level values. This is the lowest line. The re-scaling lifts the line to a new position and rotates it slightly and represents equated test logits on the vertical scale. In the same figure the individual Year 3

and Year 5 equi-percentile equating lines are shown as lines of indicated gradient. In addition the equi-percentile equating line using the fitted quadratic curve for the combined Year 3 and 5 data is shown. This line is approximately identical to the Rasch model for score conversion for large portions of the teacher profile scale. The two curves deviate below 1.3 profile units and above 4.0 units. The Year 3 and 5 lines appear to touch (or follow) the coalesced curves as approximate tangents. This indicates that the general conversion (whether Rasch or equi-percentile) over the full profile level range is sensitive to the changes in scale conversion values as the Year levels of teachers increase.

Figure 8.3 1997 English Teacher assessments – Conversion of Profile to Test Scale result: Independent Rasch model.

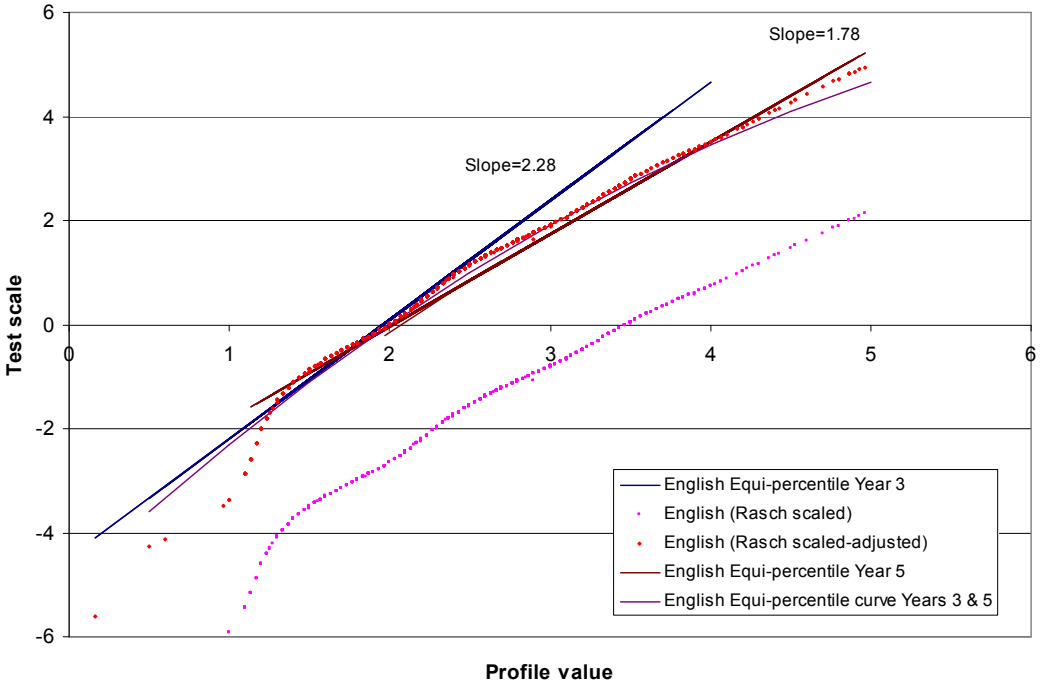
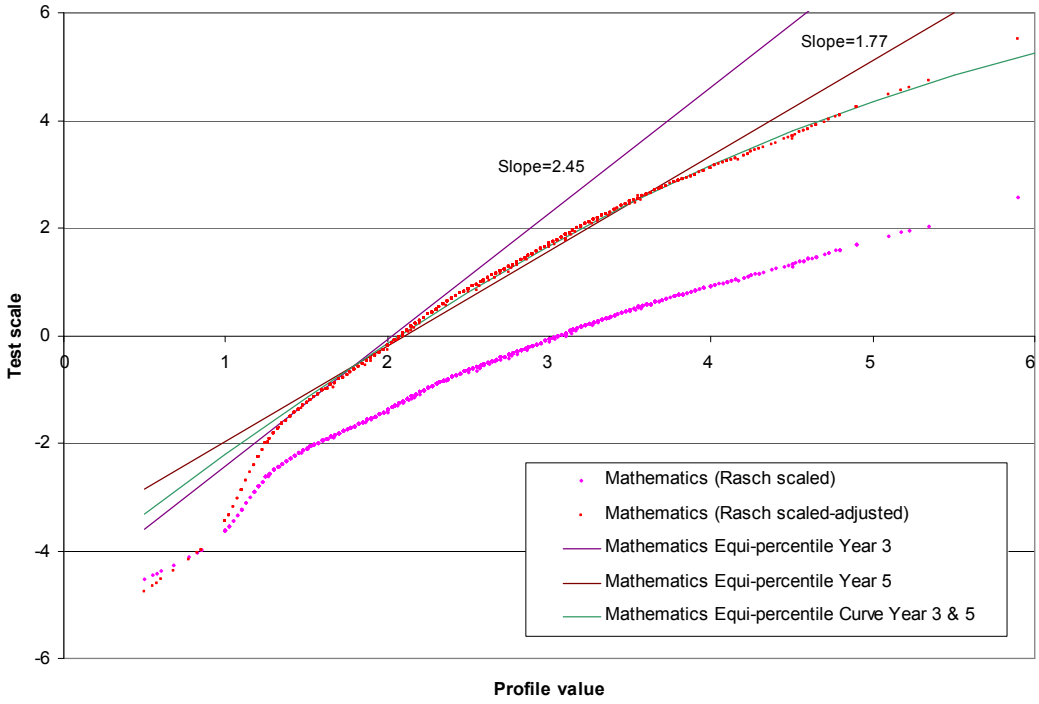


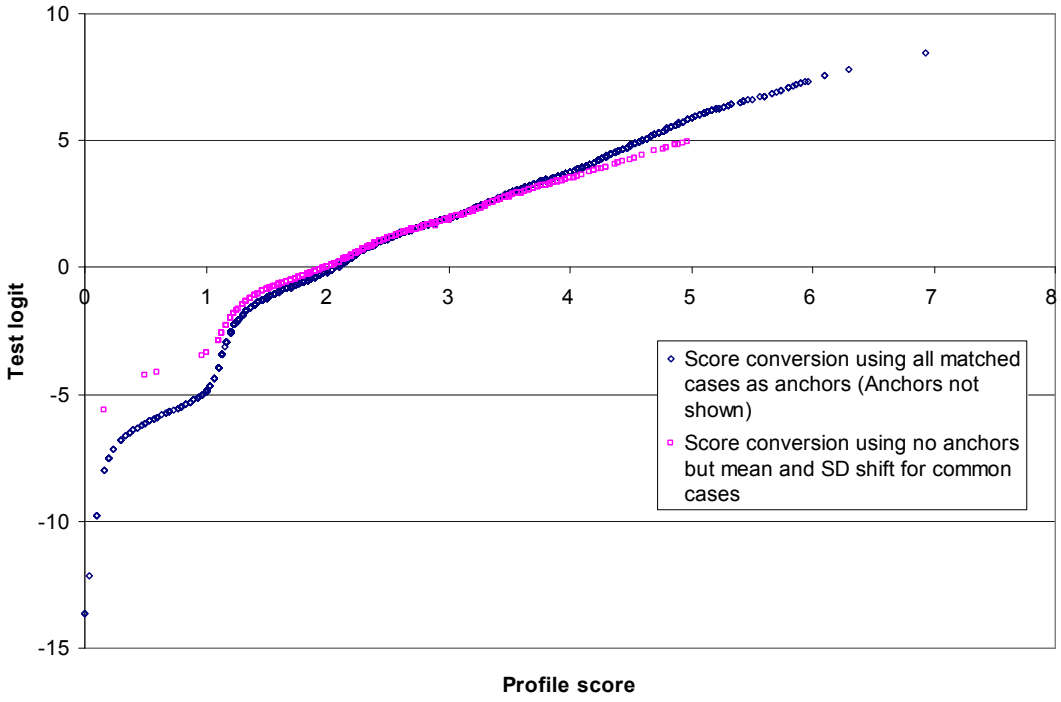
Figure 8.4 1998 Mathematics Teacher assessments – Conversion of Profile to Test Scale result: Independent Rasch model



As indicated earlier this general solution (whether Rasch or equi-percentile) averages the conversion over the profile level range, relative to the conversion relationship at each year level (the linear alternatives). As the investigation is concerned with the broad relationship only, and as the conversion can be applied generally over the full Year level range 1 to 8, the Rasch model profile level to logit conversion is adopted. That it is approximately identical to the combined Year 3 and 5 equi-percentile curve supports the adequacy of the Rasch conversion, even though the model application is a little unconventional with many over-fitting cases

Finally, in completing the initial re-scaling of teacher assessments to the scale of the test, the adequacy of the re-scaling of the teacher assessments through the Rasch model linear equating is confirmed by comparison with independently applied anchored equating. Figure 8.5 compares the linearly re-scaled result with the result of using all the common cases as anchors but suppressing the plotting of the common case points. Since the logit values obtained for the anchored approach are already directly linked to the test scale values no additional re-scaling is required. The conversion lines are very similar, as above, in the range 1.0 to 4.0 on the profile level scale.

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



Comparing Teacher and Test Assessments for Common Students with Teacher Assessments Re-scaled.

1997

The 1275 cases where students have both a test assessment and a teacher assessment, now converted to the test scale, can be compared. Each student with a test and teacher assessment in test logits also has error of measurement estimates for both assessments. Using the approach described by Wright and Stone (1999) for items and Bond and Fox (2007) for persons, the control lines for 95% confidence for each assessment on the two scales can be applied to the scatter plot of data points shown in Figure 8.6 in order to examine the invariance of the person measures across assessment types.

For a 95% confidence range, control lines are set so that the perpendicular distance of the control line from the 45 degree identity line is 2T, where T is the unit of error related to the difference between the two measures for the case. The standard unit of error of the difference calculated on either axis for person$_i$ is

$$S_{12i} = (s_{1i}^2 + s_{2i}^2)^{1/2}$$

with $T_{12i} = \left[(s_{1i}^2 + s_{2i}^2) / 2 \right]^{1/2} = S_{12i} / \sqrt{2}$.

The upper control line can be plotted with the coordinates

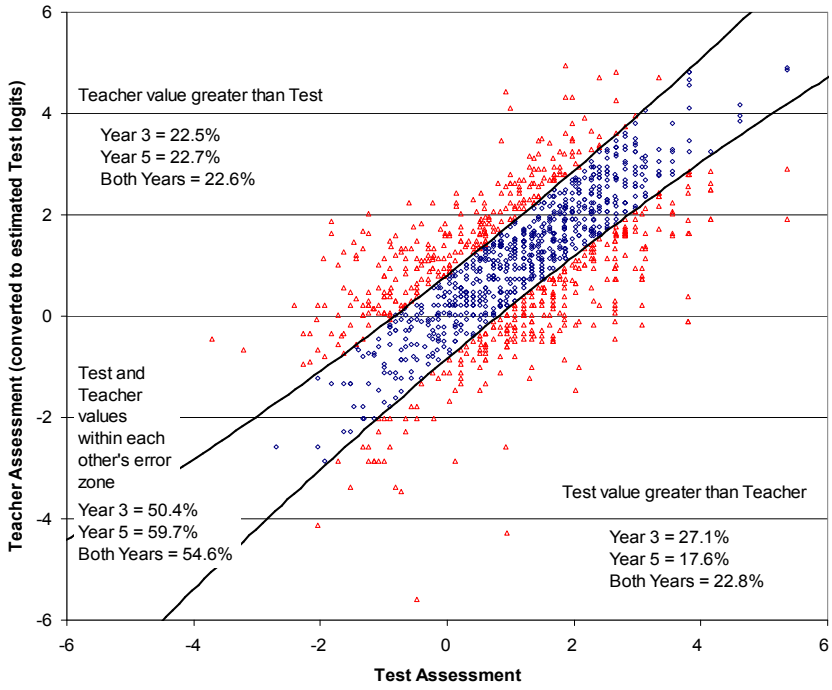
$$X = d - 2S_{12}/2 = (d_1 + d_2 - 2S_{12})/2$$
$$Y = d + 2S_{12}/2 = (d_1 + d_2 - 2S_{12})/2$$

where $d = (d_1 + d_2)/2$, d_1 and d_2 being the values of the measures on each axis, and S_1 and S_2 being the standard errors of measurement. (Based on Wright & Stone, 1999, p. 71).

The lower control line is symmetrical and plotted by reversing the X and Y coordinates.

The resultant (idealised) control lines are shown in Figure 8.6. The number of cases that can be regarded as matched within a confidence range of 95% of the errors of measurement, fall within the boundaries of the control lines, using the rescaled error of measurement values for the teacher measure. The estimates of the proportions of cases that can be considered as equivalent on both measures for English/Literacy are shown in Table 8.4.

Figure 8.6 1997 English/Literacy - Scatterplot of Teacher assessment and Test assessment invariance



The overall match rate of the two assessment processes of the combined Year 3 and Year 5 data is 54.6% of the cases. The proportions of cases above the upper control line and below the lower control line are equivalent, indicating that it is as common overall for the teacher score to be above the test score as it is for the test score to be above the teacher score when the scores do not match.

Table 8.4 1997 Comparison of Teacher and Test assessments of common students

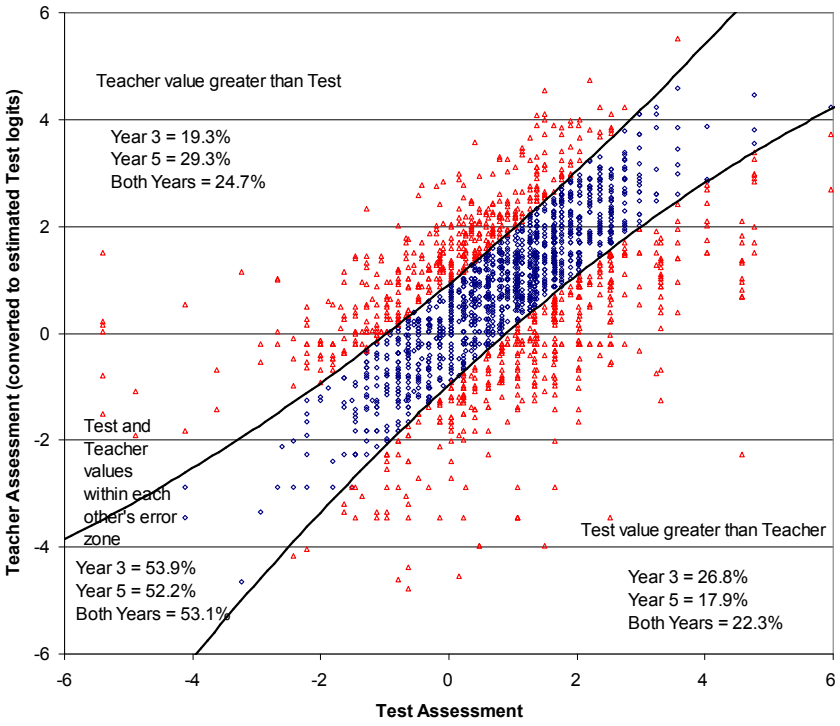
	Year 3	Year 5	Both Years combined
Teacher assesses student above Test	22.5%	22.7%	22.6%
Test assess student above Teacher	27.1%	17.6%	22.8%
Both processes within error zone of each other	50.4%	59.7%	54.6%

When the cases are examined for the specific Year levels the match rates vary. More Year 5 assessments match than do Year 3 assessments. (Part of the reason for this relates to the greater closeness of the original Year 5 only conversion scale to the combined conversion scale.) The rate of teachers assessing significantly above the test score is constant by Year level at 23% rounded. At Year 3 the test provides a score higher than does the teacher in 27% of cases. At Year 5 this test above teacher score rate is a lower 18%, with an increased match rate to 60%. The other possible sources of assessment error in the test process and the teacher assessment process, beyond the measurement error estimated within the Rasch model, are considered later.

1998

The situation for the 1998 data is summarised in Figure 8.7 and Table 8.5. The rescaling of the teacher assessments to the scale of the test required the size of the teacher assessment logit to be increase to a greater extent than for the 1997 data. As a consequence the teacher error estimates reflect proportionately adjusted values to match the test scale logit. The scatter plot patterns for Mathematics/Numeracy are similar to those for English /Literacy. The proportion of cases that fall within the control lines is 53%, slightly less than for English/Literacy. As for English/Literacy the cases outside the control lines are balanced at about 22-24%. For the individual Year levels the cases within the control lines remain about at 52-54%.

Figure 8.7 1998 Mathematics/Numeracy - Scatterplot of Teacher Assessment and Test assessment invariance



In mild contrast to the English/Literacy situation the proportions of teachers assessing the student more highly than did the test are greater for Year 5 than for Year 3. As a corollary, the proportion of test scores higher than the teacher is greater for Year 3. The equating of teacher and test scores using the Rasch model approach has averaged out the steeper conversion line of teacher assessments to the test scale that applies to Year 3 in isolation. If the steeper Year 3 conversion line were used it should make no difference to these proportions, as it can be seen from Figure 8.4 the conversion line follows the Year 3 linear gradient for most of the range in which Year 3 assessments are placed.

Table 8.5 1998 Comparison of Teacher and Test assessments of common students

	Year 3	Year 5	Both Years combined
Teacher assesses student above Test	19.3%	29.3%	24.7%
Test assess student above Teacher	26.8%	17.9%	22.3%
Both processes within error zone of each other	53.9%	52.2%	53.1%

Summary of rates of teacher assessments matching test assessments

Accepting the assumptions and results of the Rasch model equating process, teachers' assessments match test assessments in just over 50% of the cases, allowing for errors of measurement. This degree of matching occurs in two independent sets of assessments one year apart. By categorising the scale on both the test and teacher axes into 1 logit categories,

a Cohen's Kappa value of just above 0.4 is obtained. This is regarded as a fair to moderate agreement only (Altman, 1991). The combining of well-calibrated, moderately-calibrated and poorly-calibrated teachers into the one analysis leads to this relatively low agreement rate. Later in the chapter it is shown that at some school sites much higher agreement rates apply.

The spread of the error zone reinforces that all assessments are made with error. One element, the modelled error, is used to set the control lines. The modelled error, the estimate of the range within which the actual score might lie when the assessment process fits the Rasch model, is quite large. The general size of mean measurement error for the tests in 1997 is between 0.33 to 0.37 test logits (Table 6.2) and 0.27 for the teacher assessments (Table A11.2). The 1998 equivalents are 0.39 to 0.49 for the tests (Table 6.3) and 0.28 for the teacher assessment (Table A11.6). An error of 0.3 test logits is equivalent to about 7 to 9 months of learning development based on Hungi (2003), and is a direct consequence of the relatively small number of items routinely used in testing situations.

Part of the complexity in making the comparison of teacher and test assessments is the use of a single standard procedure, assumed to work consistently in all applications (the test) with a looser but still standardised process open to more varied application and interpretation (the teachers). Assuming that the processes are assessing the same learning trait, more replications of the assessments for individual teachers would provide the data to tease out the potential sources and causes of disagreement in the assessments. The current data sets cannot offer much insight into the likely reasons when assessments do not match. However, within the error tolerances of the assessment processes and the model for equating, slightly more than 50% of the students can be regarded as having invariant assessments across forms. As will be revealed later, there is evidence that at some individual school sites (as proxies for teachers) the number of cases that match is low but the correlation between the two assessment processes is high. This indicates that considering the matching alone is an inadequate basis for comparison.

There are very few published examples of the relative performance of teachers and tests in assessing students using common scales. Examples in Chapter 4, in particular Tables 4.2 to 4.4 and Figure 4.3, indicate degrees of match but in most cases using a much less precise scale The broader the scale unit, the greater the chances of teacher and test assessments matching. An estimated match rate at Key Stage 3 of 61.5% for English and 70% for mathematics (Figure 4.3) uses the very broad unit of one Key Stage level. Teachers in England were also assessing to a more explicit framework; which was also used by test designers to develop the tests. One of the sources of variability is the teacher. Can the data provide an insight into the effect of teacher assessment skills on the extent of match?

Estimates of between teacher differences in matching

As indicated earlier there is no way that individual teachers in the data set can be identified. By design the students assessed by particular teachers cannot be grouped together. However as a consequence of the collection process it is possible to aggregate data at each school site, after recoding site codes to remove their specific identities. Teachers provided, on average, 5 student assessments each. For sites with more than 5 students per Year level, the site data are for multiple teachers. Thus the teacher-assessment test relationships for small groups of teachers at each site can be established. The general match rates in the previous sections can then be re-examined within and across sites.

Figures 8.8 and 8.9 show the distributions of the match rates (i.e., invariance within measurement error) based on sites converted to the estimated numbers of teachers at each site. The mean match rate at the site is ascribed to all teachers. In practice teachers at a site would also be likely to vary in their match rates.

Figure 8.8 Match rates 1997 - English/Literacy

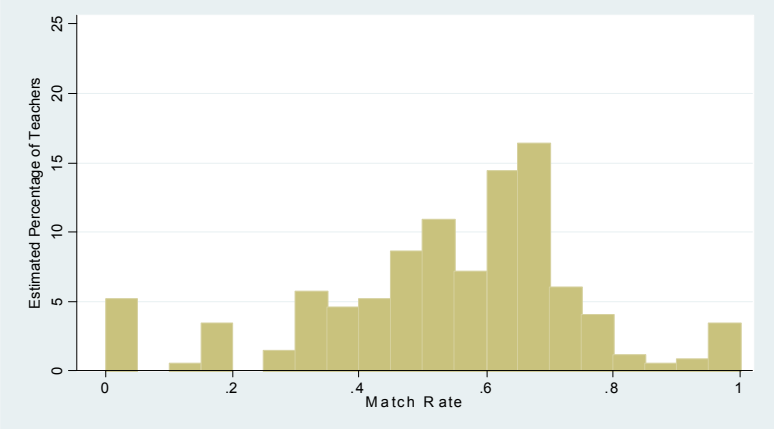
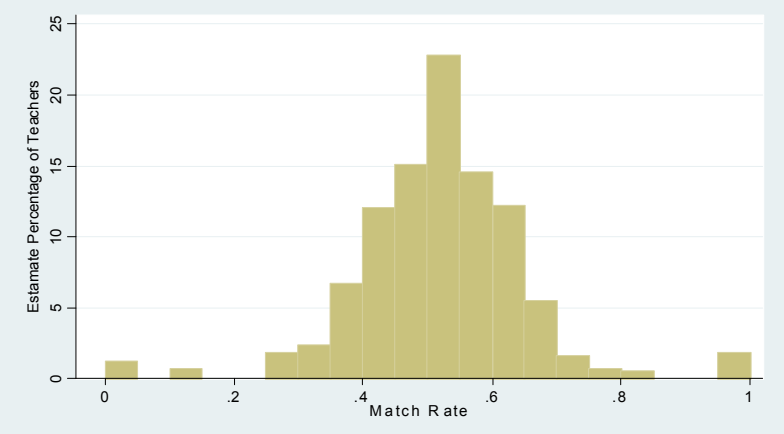


Figure 8.9 Match rates 1998 - Mathematics/Numeracy



The distributions appear to be different. This may be learning area related, although as raised earlier there is the possibility of non-random loss of data as a result of the process to connect

test and teacher judgement assessment for individual students. The English/Literacy distribution is less concentrated around the mean match rates, with proportionately more cases with no match and with high match. Although both learning areas have similar overall match rates, the mode values differ (near 0.7 for English/Literacy, nearer 0.5 for Mathematics/Numeracy). It appears that there is greater variability in the match rate of assessments in English/Literacy than in Mathematics/Numeracy. Some English/Literacy teachers are well aligned to the test scale (above say a match of 0.7). Fewer Mathematics/Literacy observations were as well matched. However some English/Literacy assessments are very poorly aligned (below a match rate of 0.3) relative to Mathematics/Numeracy.

Six case studies for English in Figure 8.10 and Table 8.6 illustrate an approximation of what the situation might look like when multiple students for individual teachers are examined. The cases are selected on the basis of relatively high numbers of students at a site (above 13 implying at least three teachers) and for a range of match rates. Match rates are calculated as the proportion of measurably invariant assessments relative to the total number of assessments. The highest match rate of the selected cases (site 1687) is 0.92, the lowest 0.16 (site 2777). The scatter plots all have positive slopes and positive correlation coefficients. For site 1222 there are some outlying cases. Sites 1936 and 2777 have quite varied matching rates (0.7 and 0.16) but high correlation coefficients (0.95 and 0.79). These cases highlight the matters raised in Chapter 4 on forms of matching. Intercepts and gradients using TLS/Deming regression advocated in Chapter 4 (as distinct from OLS) are included as broad indicators of the variation in teacher assessment matches with the test score across sites.

Figure 8.10 1997 English/Literacy: Comparison of Teacher assessments (Rasch model equated) and Test Model assessments at selected sites.

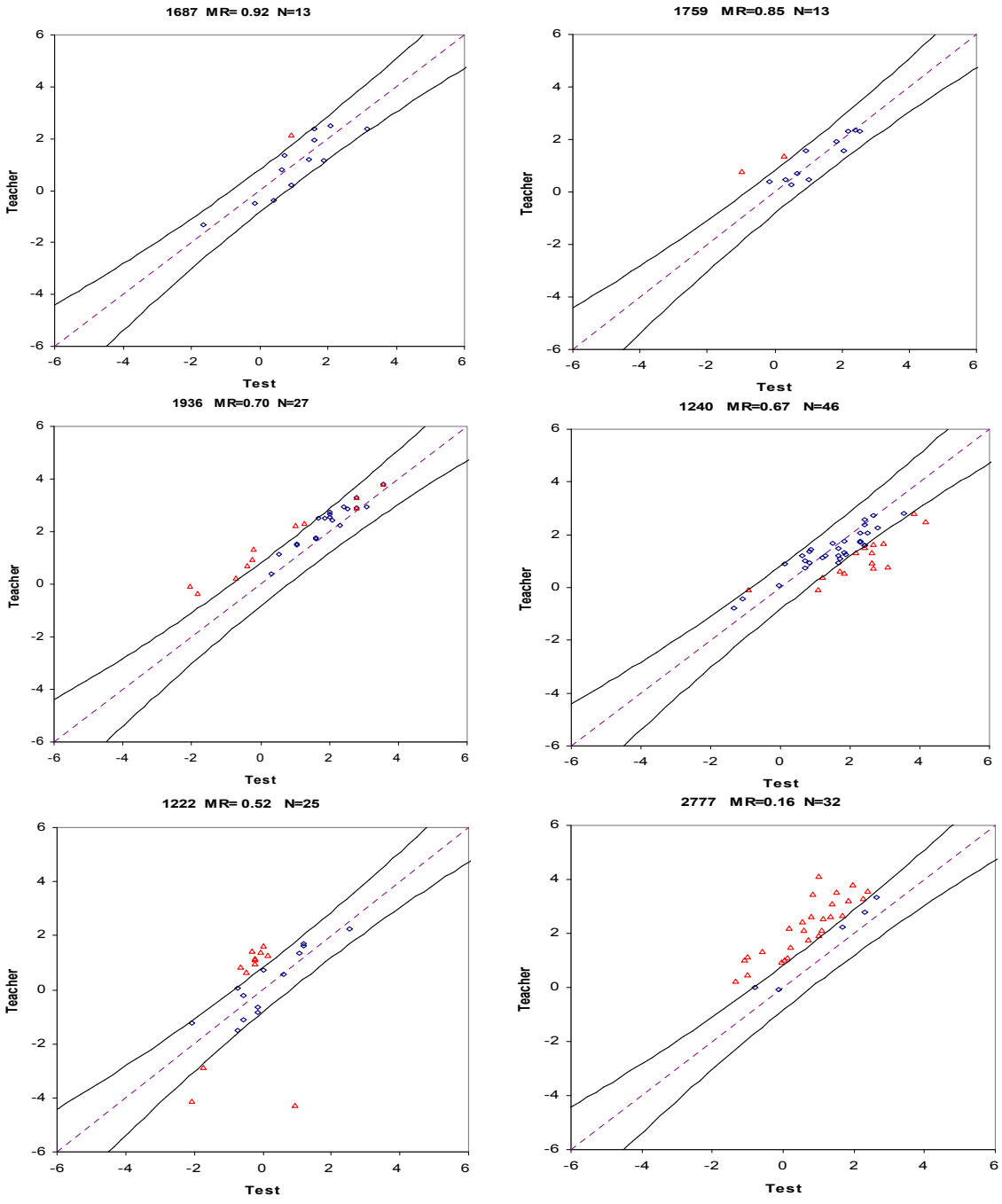


Table 8.6 1997 "Deming" Regression and Kappa values for Teacher and Test assessments of common students at selected sites

Site	Test Mean	Teacher Mean	Sample size	Corre-lation	Match Rate	Variance ratio	Intercept	Std. Error	Slope	Std. Error	Kappa
1687	1.05	1.07	13	0.85	0.92	0.96	-0.07	0.34	1.08	0.27	0.45
1759	1.04	1.28	13	0.78	0.85	0.67	0.59	0.32	0.65	0.19	0.50
1936	1.27	1.92	27	0.95	0.70	0.44	1.00	0.11	0.73	0.05	0.58
1240	1.75	1.30	46	0.84	0.67	1.82	0.11	0.13	0.68	0.06	0.54
1222	-0.13	0.07	25	0.54	0.52	2.87	0.44	0.56	2.93	1.47	0.41
2777	0.73	2.13	32	0.79	0.16	0.12	1.47	0.13	0.91	0.09	<0

The assessments in the lower panels have fewer cases within the control lines but have medium to high correlation coefficients. Deming regression analysis (allowing error on both axes) indicates some reasons for mismatch related to teacher calibration to the test scale.

At site 2777, with a low match rate of 0.16, many cases are above the upper control line and thus do not meet the criteria for a high match of assessments. The correlation is high, the slope for the regression close to 1 (0.91 with low standard error, 0.09) and with an intercept on the teacher axis at 1.46 (standard error of 0.13). Taken together these data points imply a high degree of calibration to the test scale, but with teacher assessments consistently of the order of 1.5 logits above the test. From the slope (0.91) it is seen that the scale range for teachers is slightly narrower than the test scale. The teachers are however clearly following the scale of the test but their assessments are displaced consistently above it.

Kappa values are obtained by categorising the assessments into 1 logit wide categories on each scale. Apart from site 2777, positive values above 0.4 are obtained, indicative of a fair to moderate agreement (Altman, 1991). For site 2777, the Kappa value of less than 0 implies a lower than chance match. If however the scale categories for the Kappa calculation are re-categorised after a 1.5 logit shift down on the teacher scale, the Kappa value becomes 0.58, equivalent to the highest Kappa in Table 8.6. A major reason for the assessments not matching at this site is the teachers systematically assigning higher values to students relative to the test assessments, leading, to an over-estimation of scale positions by teachers relative to the test. Rescaling of all cases down by 1.5 logits leads to a match in most cases, implying these teachers are calibrated to the scale but systematically over estimate scale positions.

Most of the case study sites show consistency in assessments within a school, even though the test and teacher assessments may not be measurably invariant. The correlation coefficients are at or above 0.78, except for site 1222. This puts the selected case studies mostly above the overall correlation coefficient for the full 1275 cases of 0.66. This reinforces that these are selected sites (on the basis of the varying match rate across the matching scale and relatively high correlation coefficients) and thus do not necessarily represent the general pattern. However the scatter plots suggest it is possible to have multiple teachers at a site (*n* estimated to be 6 for site 2777) assess consistently at Years 3 and 5. That is they all seem to follow the same general understanding of learning status even though this common understanding is systematically displaced from the test calibration. This observation can be made of all exemplar sites except 1222 where more outliers indicate greater variation in teacher and test perspectives.

In the case of site 1222 the Deming regression indicates through the high slope value (2.92) that teachers have a markedly wider scale range for their assessments than does the test. The

reverse applies for site 1240; a narrower teacher than test range. Both examples illustrate that the order of students can be approximately consistent for the teachers and the tests but without a calibration process to ensure that the scales are seen to have equivalent units, the usefulness of the teacher chosen scale value, as a description of learning status, is diminished. Both examples, when seen relative to the other examples, offer hope that it is feasible to attempt to train teachers to locate their perceptions of learning development on a common scale.

Table 8.6 provides the Kappa value for strength of agreement between the two assessment methods at each site. Assessments are categorised into categories 1 logit wide on each axis to calculate Kappa. Based on Altman (1991) most agreements are either moderate (0.41-0.60) or fair (0.21-0.4). For case 2777 as explained above, adjusting each teacher assessment downwards by 1.5 logits leads to a revised Kappa of 0.58, confirming that the main reason for assessments mismatching is the systematic misalignment of teachers to the test scale. When a statistical adjustment is made the match rate becomes 0.94, only two cases remain outside the confidence limits. The case studies illustrate that for a number of teachers in English/Literacy the test and teacher scales are closely related even when the match rate is low. At some sites the scales correlate less well. The potential is there however to study those teachers who align well, to attempt to understand and develop processes to train other teachers to be so aligned to the test scale.

Figure 8.11 1998 Mathematics/Numeracy: Comparison of Teacher assessments (Rasch model equated) and Test Model assessments at selected sites

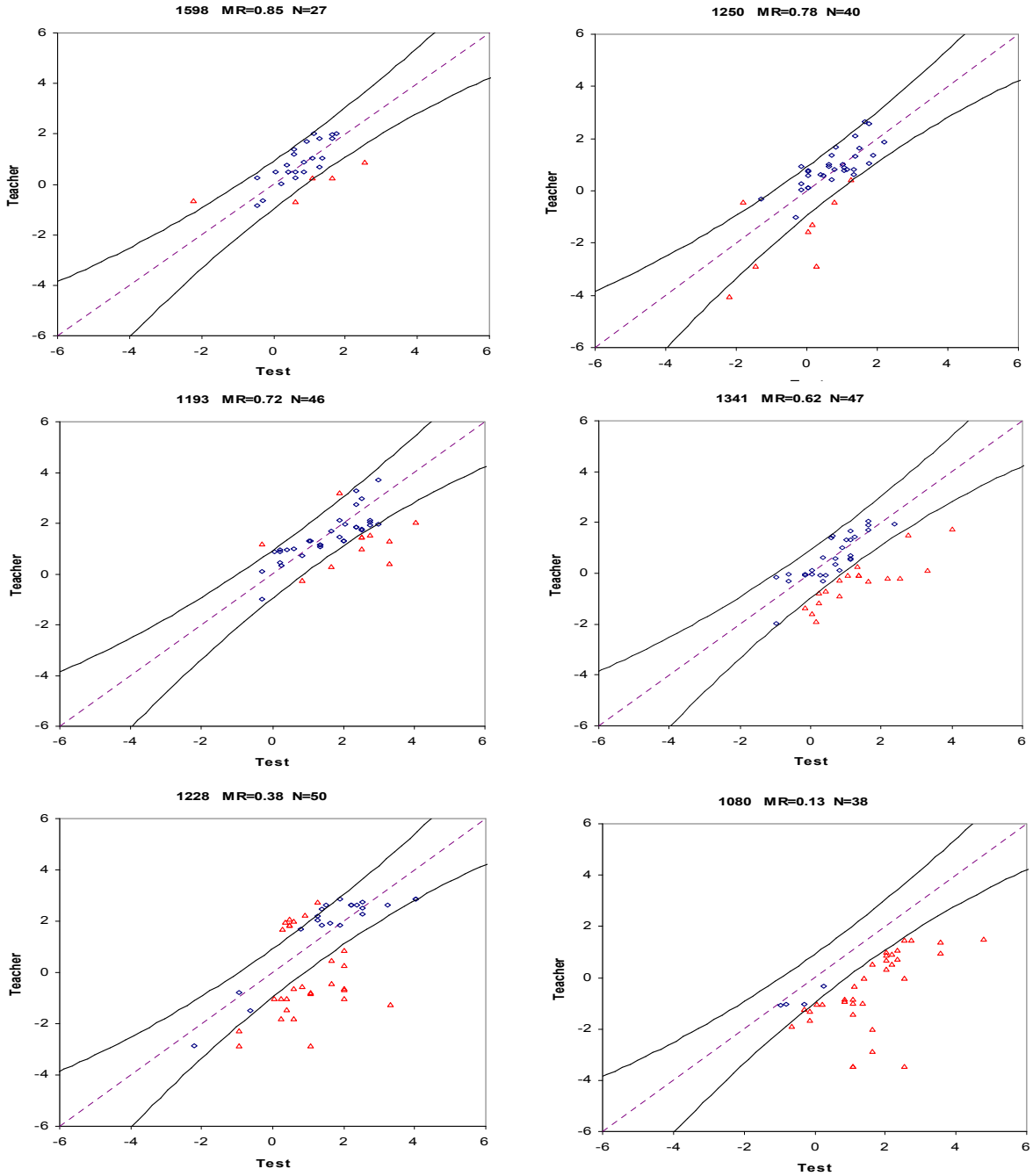


Table 8.7 1998 "Deming" Regression and Kappa values for Teacher and Test assessments of common students at selected sites

Site	Test Mean	Teacher Mean	Sample size	Corre-lation	Match Rate	Variance ratio	Intercept	Std. Error	Slope	Std. Error	Kappa
1598	0.72	0.72	27	0.65	0.85	1.00	0.08	0.37	0.89	0.39	0.41
1250	0.52	0.40	40	0.75	0.78	1.66	-0.46	0.28	1.66	0.26	0.42
1193	1.81	1.47	46	0.68	0.72	1.53	0.12	0.27	0.74	0.16	0.36
1341	0.89	0.25	47	0.54	0.62	12.71	-1.30	0.37	1.74	0.34	0.37
1228	1.22	0.66	50	0.57	0.38	3.41	-2.25	0.85	2.38	0.58	<0
1080	1.34	-0.52	38	0.56	0.13	7.07	-3.02	0.97	1.80	0.60	<0

Figure 8.11 and Table 8.7 show site case studies for the Mathematics/Numeracy assessments. Site 1598 represents 27 students assessed by an estimated 5 teachers. All but 4 assessment results are invariant. The Deming regression slope is close to 1 (0.89) and the intercept very close to 0 (0.08). These teachers could be regarded as approximately calibrated to the test scale. The correlation (0.65) indicates that further training might be required to improve the linking of teachers to the test scale along with examination of aberrant cases to clarify which assessment process is the less accurate.

As for English/Literacy, a low match rate does not imply a low correlation. Site 1080 with 38 cases (estimated 7 teachers) has a low match rate of 0.13 but a correlation coefficient of 0.56. The assessments are generally below the identity line, with most cases below the lower control line. Some of the outliers would suggest a poor relationship to the test scale. However ignoring the worst four outliers (particularly if they were to belong to just one teacher) produces a scatter that indicates a site consistency at least. All teachers at this site have somehow developed a consistent view among themselves of the use of the teacher assessment scale, and thus all appear to under-estimate their students learning development when the test scale is adopted as the standard.

The absence of identified individual teacher cases (and too few assessments per teacher even if they were identified) means that the judgement consistency of individual teachers cannot be observed. Analysis at a site level provides a deeper appreciation of the possibilities for teachers to become calibrated to the scale of appropriate test measures on common dimensions of learning. The cases studies are not necessarily representative of all sites but illustrate that much deeper understandings of the assessment behaviour of teachers are obtained when a site view is taken. An individual teacher view should be even more informative. Even though a site may have few assessment cases that are invariant (within error), there are sites where the teachers appear to be assessing consistently and bear a common - but displaced relationship - to the test scale. Such behaviour if confirmed elsewhere would provide a basis for building a common scale approach to student developmental assessment where teacher and test assessments could be constructively blended to provided an integrated approach to classroom assessment.

The conversion of teacher to test scales in this study is normative. It assumes the mean practice of the teachers indicates where the test and teacher scales should equate. An alternative analysis focused on specific skills and behaviours might establish a more appropriate relationship of test scores to a teacher scale or vice versa. Alternative processes to set the linkages (Hattie & Brown, 2003) might then provide a criterion basis for linking the scales. Such equating would then enable studies to establish (say in Victoria) a better

indication of the extent to which individual teachers are directly calibrated to, or consistently displaced from, the test scale allowing then the potential for individual teacher re-alignment.

Having established a scale linking process through the common cases at Years 3 and 5 it is possible to explore (speculatively) the assessments in the full range from Year 1 to 8. The next section addresses this more global comparison.

Extending the comparison of Teacher and Test assessments beyond Years 3 to 5

Comparing Teacher assessments to the Years 1 to 8 Test data model.

Using the common cases as a basis, the full set of teacher assessments is converted to estimated test logits. This conversion is made by applying the rule established in the common cases to the full range of teacher judgement assessments. The teacher data are thus expressed in test logit values rather than in profile level units for each student. These re-scaled teacher assessments can be compared with the test model developed in Chapter 6, an estimate of what the test data might look like, based on the best estimates of the trajectories of growth with age/Year level.

Figure 8.12 1997 English/Literacy-Mean of Teacher assessments (Rasch model equated) and Test Model assessments compared, by gender and Year level

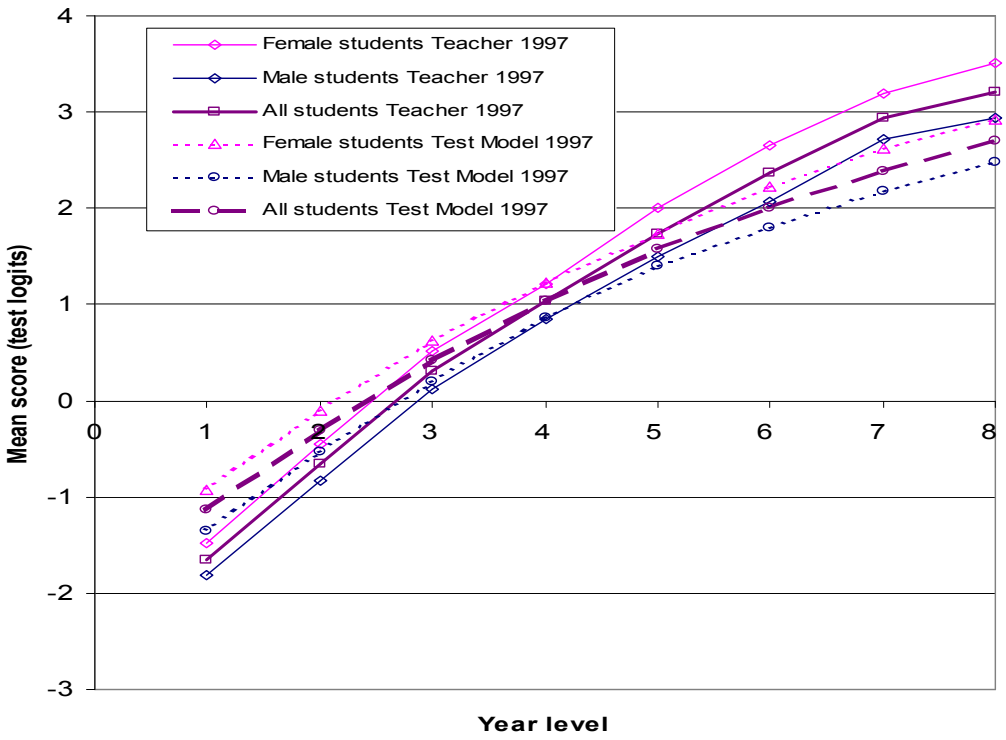


Figure 8.12 compares the two assessment processes for English/Literacy by Year level and gender. It was established in Chapters 6 and 7 that there were morphological similarities in summaries by Year level and gender for teacher assessments of English when compared with

the Literacy test. At Year 4, the notional average of Year 3 and 5, and thus at the point where the teacher-assessed cases are equated to the test scale, all students, male students and female students coincide for both assessment processes. Thus while the equating is performed on an 'all-students' basis, the gender patterns from both assessment processes are very similar (and quite different from the pattern for mathematics shown later). This provides evidence that the teacher assessments describe the Year 4 population in the same way by gender as do tests.The relative gender relationships apply from Year 1 to Year 8. The trajectories of the teacher assessments and the model test data however differ. It is not easy to establish the extent to which the trajectory difference is an artefact of the multiple assumptions that led to the establishments of the test model (only Years 3, 5 and 7 are actual data) and/or the process applied to convert the teacher assessments to the test scale. A later section will compare the data sets where the trajectories are also equated.

Figure 8.13 1998 Mathematics/Numeracy-Mean of Teacher assessments (Rasch model equated) and Test Model assessments compared, by gender and Year level

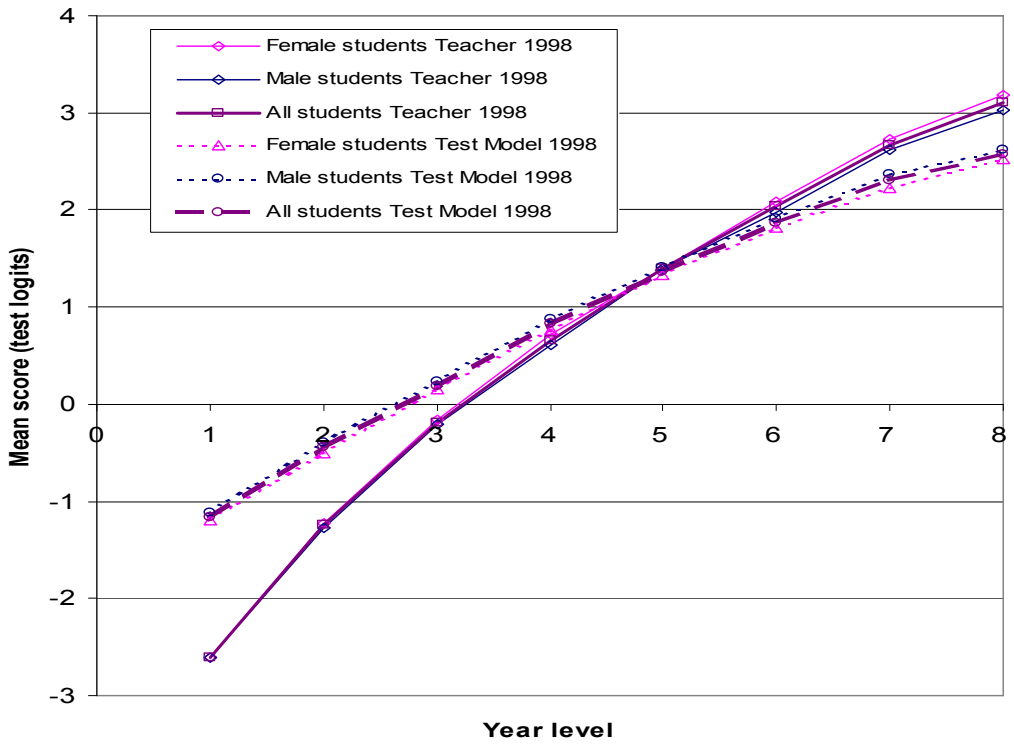


Figure 8.13 presents the Mathematics/Numeracy comparison over 8 Year levels. The general trajectory difference applies here also. There is almost no difference in the gender summaries in the two assessment processes, apart from a small reversal of the very small gender differences in the upper Year levels. The test model shows a slight advantage for males in upper years, the teacher data a slight advantage for females. The most remarkable feature is the approximate consistency in the gender view, especially when contrasted with the

English/Literacy equivalent. Teachers and test summaries show the same general pattern even though gender is not relevant in the equating process.

An implication of the apparently different trajectories is that in the lower Year level teachers generally report a lower assessment value than does the test for the same student. This difference is reversed in the upper Year levels. It cannot be established from the available data whether this situation is real or an artefact. The test model is influenced by floor and ceiling effects of the individual test and these may account for some of the differences.

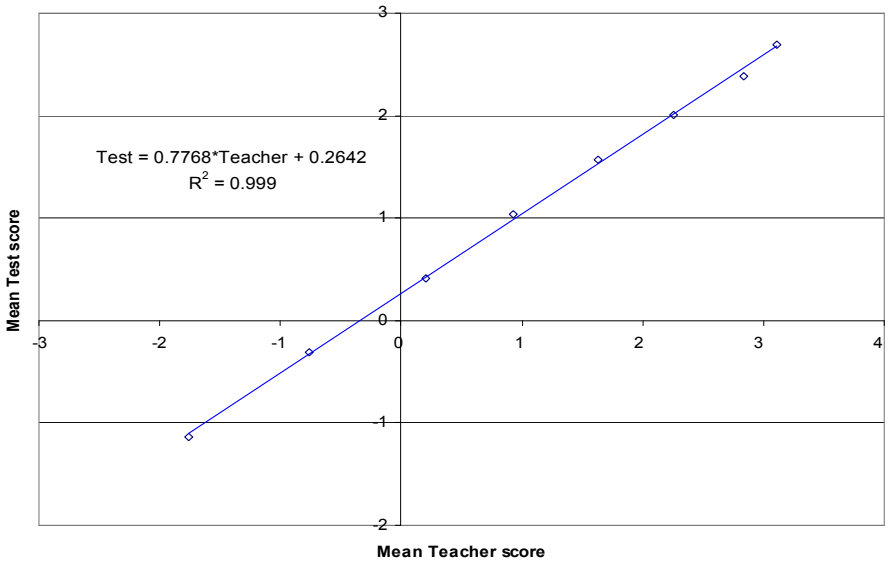
Both the English/Literacy and Mathematics/Numeracy comparisons can be presented by age within Year level. They reflect, generally, the same relationship from the test- or teacher-assessed perspective. However the difference in the trajectories makes the visual comparison by Year level complex. A comparison is made later once the trajectories are equated.

Equating the trajectories

For comparison purpose, the complicating effect of the different trajectories of test and teacher assessments by Year level on the general patterns can be neutralised by equating the trajectories. This is not an equating in the sense applied earlier in the chapter but one of convenience, on the assumption that the trajectories of learning growth with age/Year level should in principle be the same, independent of the particular assessment process. As discussed above it is quite feasible that teachers could consistently under-estimate the learning status of lower Year level students and over-estimate upper Year level students as reflected in the Figures 8.12 and 8.13, particularly in the absence of training and feedback. While in reality it is possible for the trajectories to be quite different, removing this aspect from the data allows a comparison of the degree to which the underlying patterns in the test and teacher assessments reflect the same general phenomena.

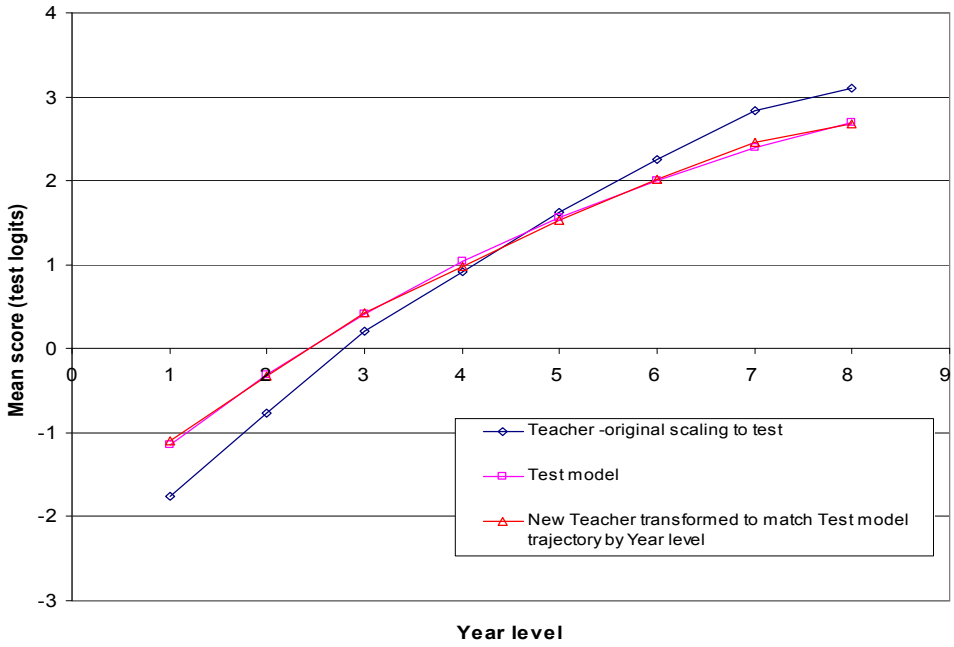
It has been established in the examination of groups of teachers at individual sites above, that at some sites teachers' assessments match the test assessments consistently, that is they are invariant within error. It also established that it is possible for teachers to be consistent assessors relative to the test scale but displaced above or below the expected norm derived relationship and to have a consistent gradient of this relationship with the test scale. The following equating process removes the effect of the difference in trajectory, even if that difference is a real effect. The trajectory equating is achieved by plotting the Year level means for the score values for the teacher and test-model scores. A line (Figure 8.14) is then fitted to the points and this used to transform teacher data (already in approximate test logit units) so that the means at each Year level are the same for both assessment processes. The choice of the test means as the base is for consistency. It does not imply that the test trajectory is the correct or real trajectory.

Figure 8.14 1997 English/Literacy Test and Teacher mean scores at each Year level-Expression to equate means



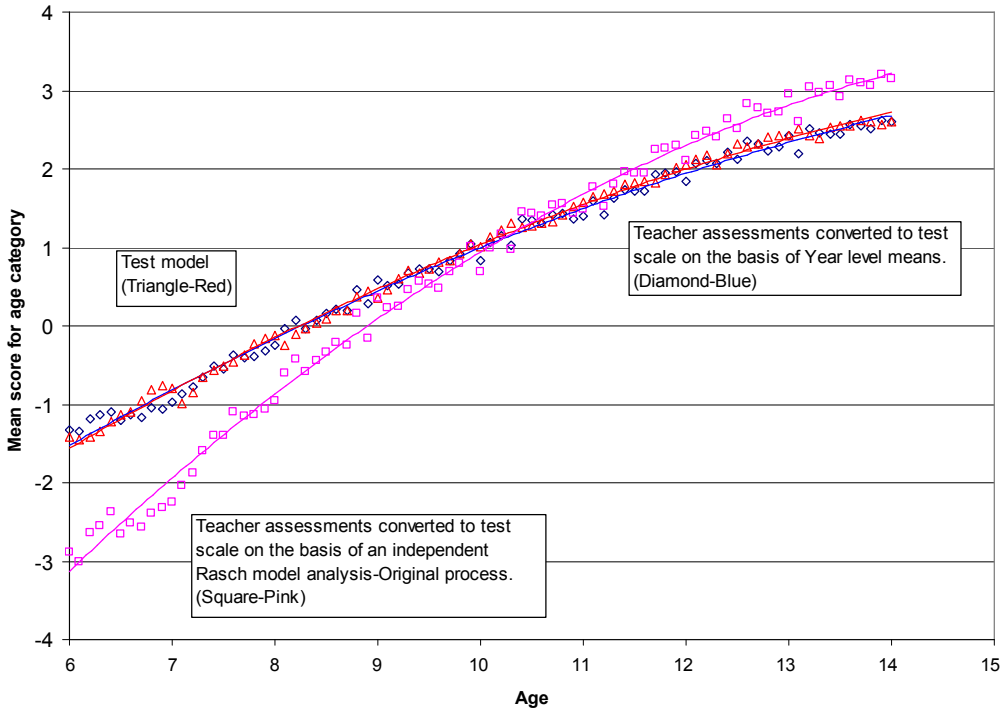
The means for each Year level are equated through linear equating. Through this process the SDs are also equated. The re-scaled teacher data are summarised by Year level to confirm the effect of the additional transformation. The original teacher assessment trajectory in Figure 8.15 is compared with the modified trajectory and the original test model trajectory. The effect of the additional re-scaling has been to make the trajectories identical as intended.

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



The result of simple mean equating by Year level of the trajectories for 1998 data, applied for simplicity in lieu of linear equating, is shown in Figure 8.16. In this case SDs are not equated. The data in this presentation are summarised by the age categories in decimal age units, reflecting that the Year level derived transformation adequately transforms the data into an age view. A quadratic line of best fit for the approximately 80 age points is applied in all three cases. The lowest curve is the original teacher data already re-scaled by the Rasch model to the test scale. The age points are transformed (to the diamond points) at the teacher assessment converted to the test scale trajectory. The interwoven curves are the test model and the mean equated teacher data.

Figure 8.16 Effect of Alternative equating processes on Teacher Test assessment comparisons-using Mathematics/Numeracy 1998



Reflection on the impact of the Year level mean transformation to the 'signals' in the data

Before summarising the data in more detail a reflection on the process is useful. Are there some steps in the process of summarising the data that have polluted the teacher or test data so that the results of strong general similarity are guaranteed? Has the development of the test model ensured that the data will match when summarised by age, year level and gender? The data sets are developed independently, at least until the equating steps. The test model is based on estimates of likely trajectories for test data as described in Chapter 6. Curve fitting to vertically scaled tests by Year level, from a number of sources, as illustrated in Chapter 5

produces a consistent generic result for many data sets. The trajectory with age and Year level is a curve with a diminishing growth rate with time. The estimate of IRT test measured student learning growth patterns, are made independently of the teacher data.

The teacher data are fitted to the Rasch model independently of the test data. The original observations by individual teachers generate the data points for each student. Nothing has been done to disrupt or change the natural teacher-observed relationships between strands, ages, Year levels or gender except through systematic transformations. The transformations are based on a Rasch analysis of 1275 (1997) and 2100 (1998) common cases at Years 3 and 5 and then the mean and SDs on the teacher scale transformed to equal those on the test for the common cases. These transformations are then applied to the full teacher data set. As far as the author can see, none of the transformations have contaminated the general trends in the data or ensured that particular relationships should be found. The analysis raises the possibility that teachers may, on average, see student development consistently with test assessments but under or over estimate a student's status depending upon the Year level or age or stage of development.

The transformation of the teacher assessment value to a logit value produces a very similar result when compared with an equi-percentile equating, suggesting that both equating process are approximately equivalent. The equating of the teacher and test scales is limited since it depends on two Year levels only out of 8 (Year 3 and 5), though these are balanced in the central zone of the Year levels of the teacher data. The limited number of common points may influence the relative Year level trajectories of the two assessment processes but will not influence other elements of the data. The effect of the difference in trajectory, even though this may be real, is removed using mean equating for each Year level. When removing the differences due to the apparently different trajectories, the transformation should not affect other general properties of the data. The transformation to equate the means at each Year level is an additional linear transformation of the teacher data. The general correlation coefficients of the teacher and test data sets with each other are unaffected.

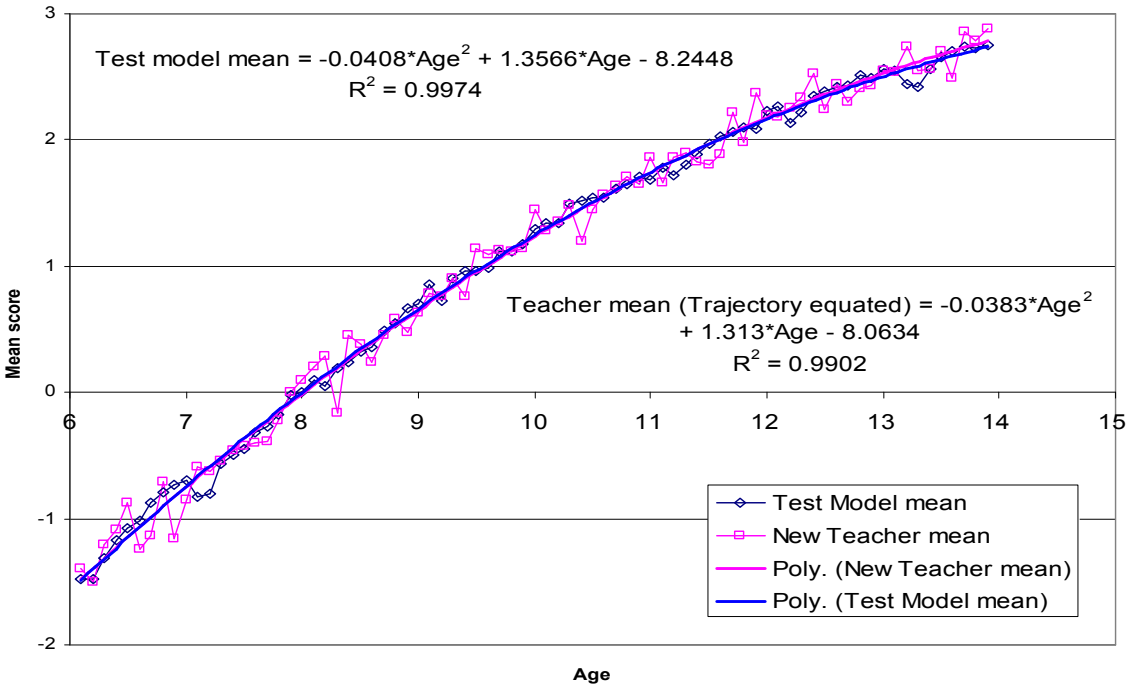
The mean equating ensures that the trajectories match. Thus this aspect, the trajectory of the relationship between test and teacher data, is artificial in the examples that follow. The subsequent comparisons merely provide confirmation that with appropriate transformations the data sets, at the level of mean summaries, can be made virtually identical. However the comparisons that are not directly trajectory dependent are the keys to understanding the degree of consistency in the two assessment approaches.

Key insights from the trajectory transformed data.

1997 English/Literacy

Figure 8.17 reinforces the view that the relationship of age and Year level is such that an equating by Year level (Figure 8.15) will also apply for age. As expected the trajectories of the means by age are virtually identical (as required by the process). In principle however two data sets, when made to follow equal trajectories based on Year level means could show more erratic relationships to the general trend. There are some points for the teacher means (pink squares as points) at each age that vary more widely from the general trajectory than do the points in the test model (blue diamonds). The mean difference from the curve for the teacher means is 0.1 logits over the age range considered. For the test model the mean difference from its curve is 0.05 logits, confirming a closer fit for the test model. The test model means are based on 64,000 cases, the teacher means on 7,900 cases. However the high R^2 values for the quadratic curves suggest that both age relationships with assessment scores are very good fits to the data. The conclusion is that both data sets are very similar. Given that the transformation was based on Year level means (not age) good fit to a similar trajectory implies inherent properties in the teacher data set that follow the same patterns with age as the test model.

Figure 8.17 1997 English-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age



The data in Figure 8.17 can be separated into gender subsets. The result is shown in Figure 8.18. The gender subsets follow essentially identical trajectories, illustrated by fitting quadratic curves. This effect is not determined by the equating of general Year level

trajectories. The teacher data could fit the general test trajectory without the gender subsets of the teacher data matching the test subsets. That the gender trajectories are very similar offers confirming evidence that teacher judgement assessments are remarkably consistent and at a population level (as distinct from individual cases) the general underlying trends and identification of learning status by gender are common to both processes.

Figure 8.18 1997 English-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by gender by age

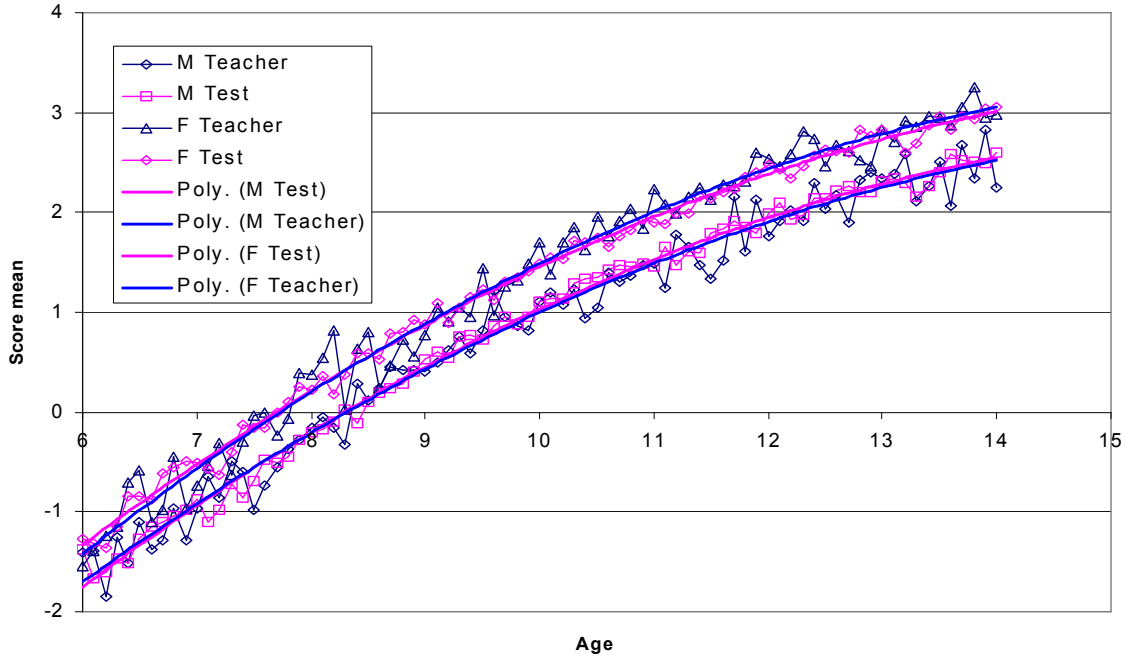
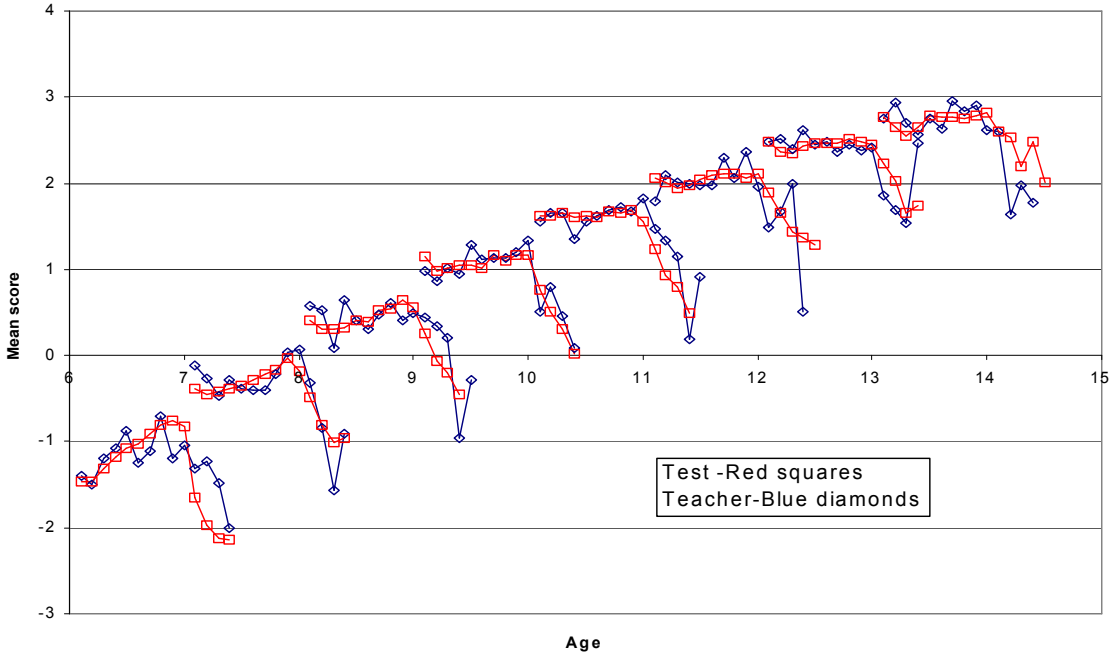
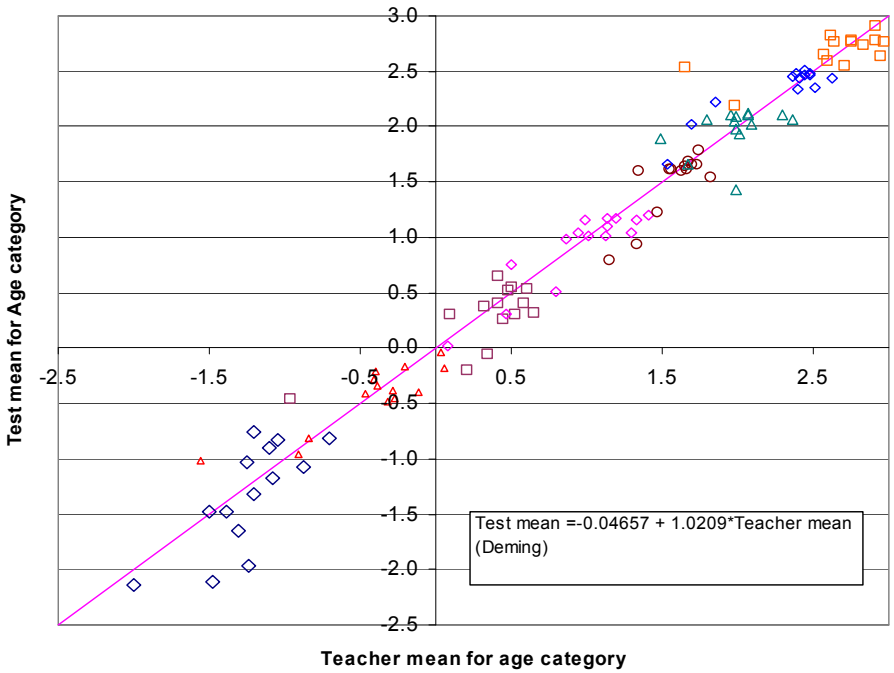


Figure 8.19 1997 English-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age within Year level



When the cases are summarised by age within Year level the relationship of the test means by age to the teacher means is indicated in Figure 8.19. That some points at each year level are coincident is to be expected as a result of the equating of trajectories. However the consistency of the proximity of many points is not a necessary consequence of the trajectory equating. In particular the trailing off of learning status estimates with age above the normal age range for the Year level appear to coincide very well, and are consistent with international data cited in Chapter 5. The age within Year level data are represented in Figure 8.20 plotted along the identity line. A Deming regression of the points results in a regression of Test mean=-0.047 + 1.02*Teacher mean (Standard errors: Intercept 0.035, Slope 0.021) indicating that the assessments are trivially displaced from the identity line and thus confirming a high consistency of assessments by age within Year level under both assessment processes. The mean assessments by age, at the very refined scale of 0.1 of a year of age, are very similar across the range of Year levels 1 to 8. Neither the test model development nor the equating processes have introduced the refined age related characteristic into the data.

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)

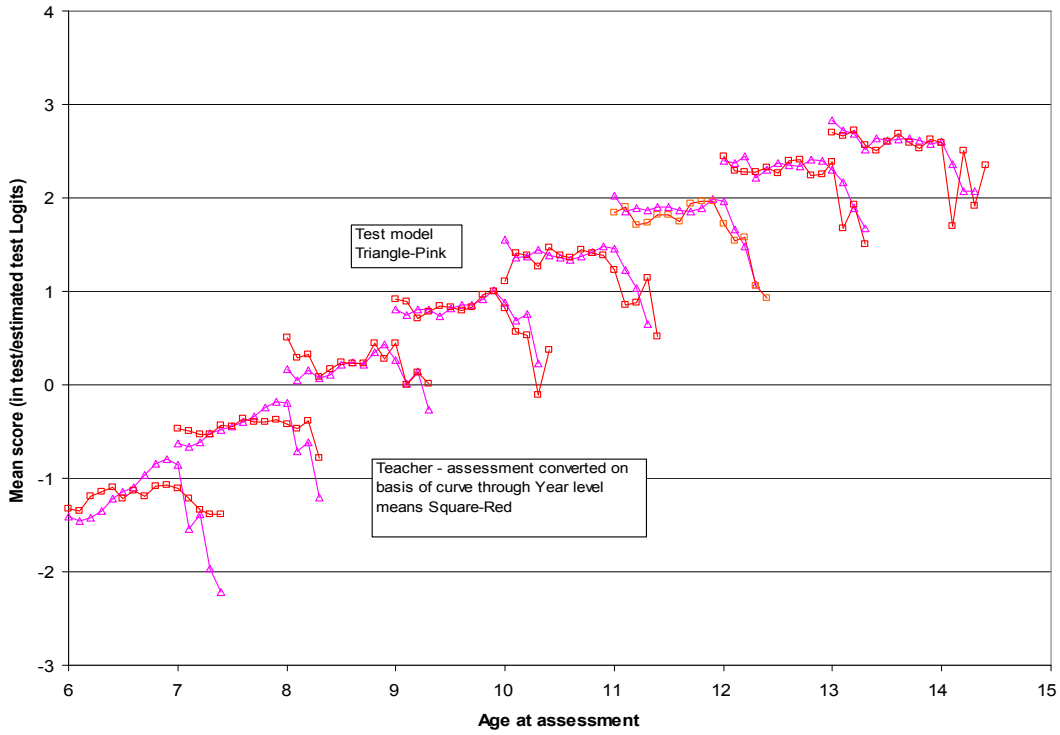


1998 Mathematics/Numeracy

The coalescing of trajectories for the test model and teacher assessments for Mathematics/Numeracy is confirmed in Figure 8.16. The equating of trajectories is on the basis of Year level means only. The same general structure of very closely coinciding data

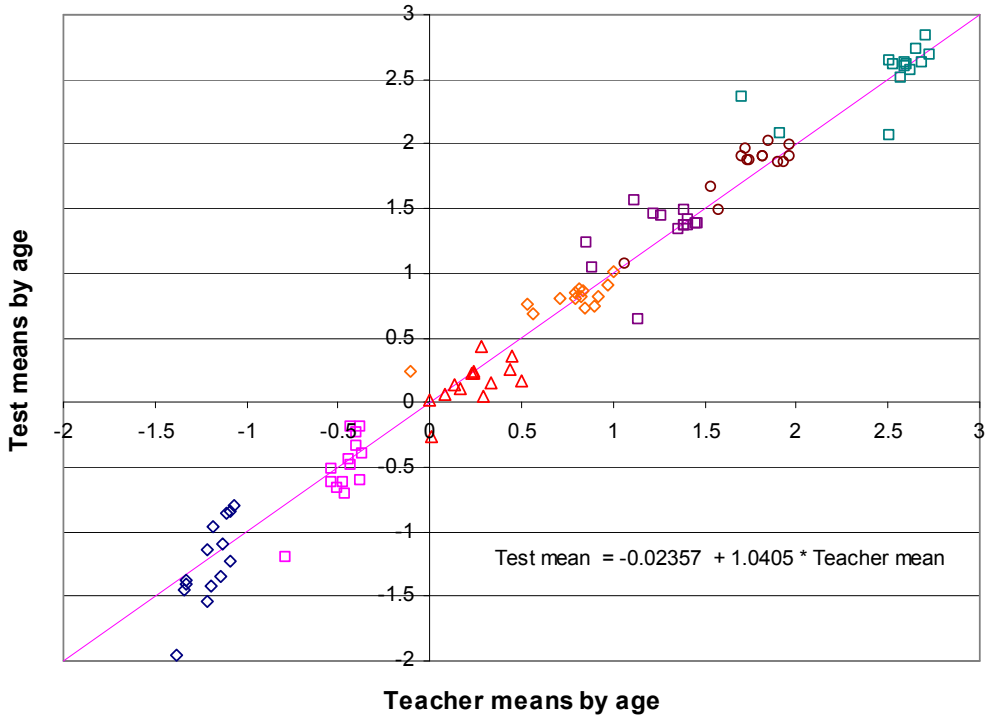
points applies for 1998 data as applied for 1997. The coincidence is less at Years 1 and 2 possibly a result of not adjusting the spread at each Year level. The flatness of the Year 1 was highlighted earlier in Chapter 7.

Figure 8.21 1998 Mathematics-Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age within Year level



The relationship of the age data points for each assessment process is illustrated in Figure 8.22. The points cluster along the identity line. A Deming regression of the points has a slope of 1.04 and an intercept of –0.0236 confirming a close fit of the points at each Year level.

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)



The plot of the gender views from a teacher and a test perspective plotted by age coincide so well that plotting the points of all four components on the one graph generates very close lines. The gender views of mathematics assessments are presented separately in Figures 8.23 and 8.24. Consistent with the English by gender view, the female and male subsets are virtually identical for both assessment processes. However for mathematics, in contrast to the English language examples, both assessment sources indicate almost no difference by gender. This illustrates that the Numeracy test assessments and the mathematics teacher assessments show little difference by gender, whereas the English/Literacy data show a strong gender differences in favour of females consistently in both assessment modes.

There is an indication that teachers in the upper Year levels see the learning status of female students slightly more favourably than do the test data (Figure 8.22), with the corollary that teacher assessments are slightly less favourable for males at the upper Year levels. Overall the two assessment processes produce very similar aggregate results for mathematics indicating that mean teacher assessments and test assessments are almost identical (when trajectories by Year level are equated).

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

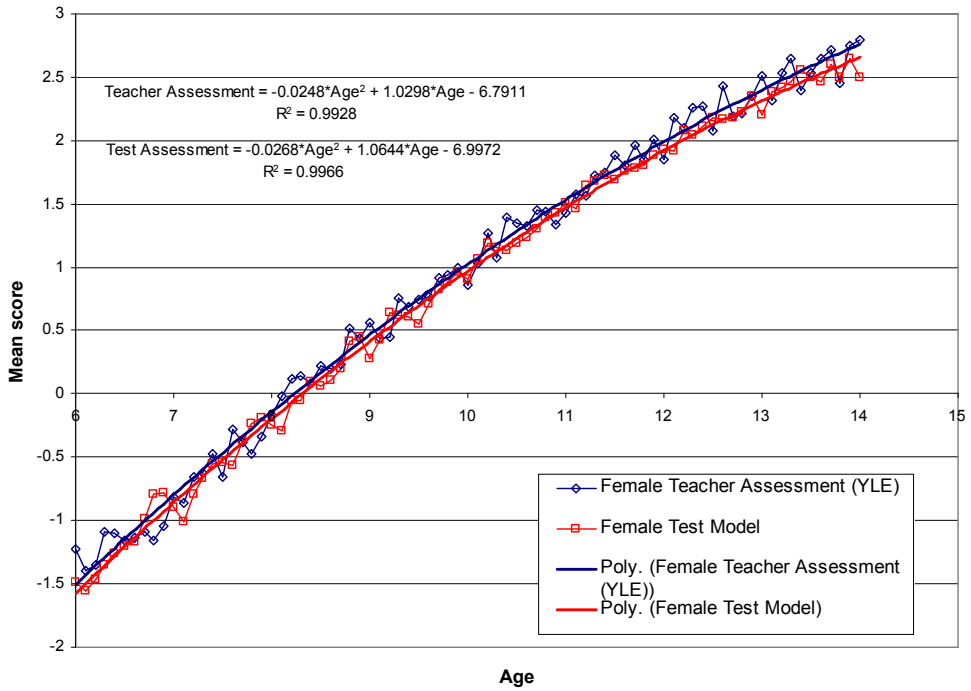
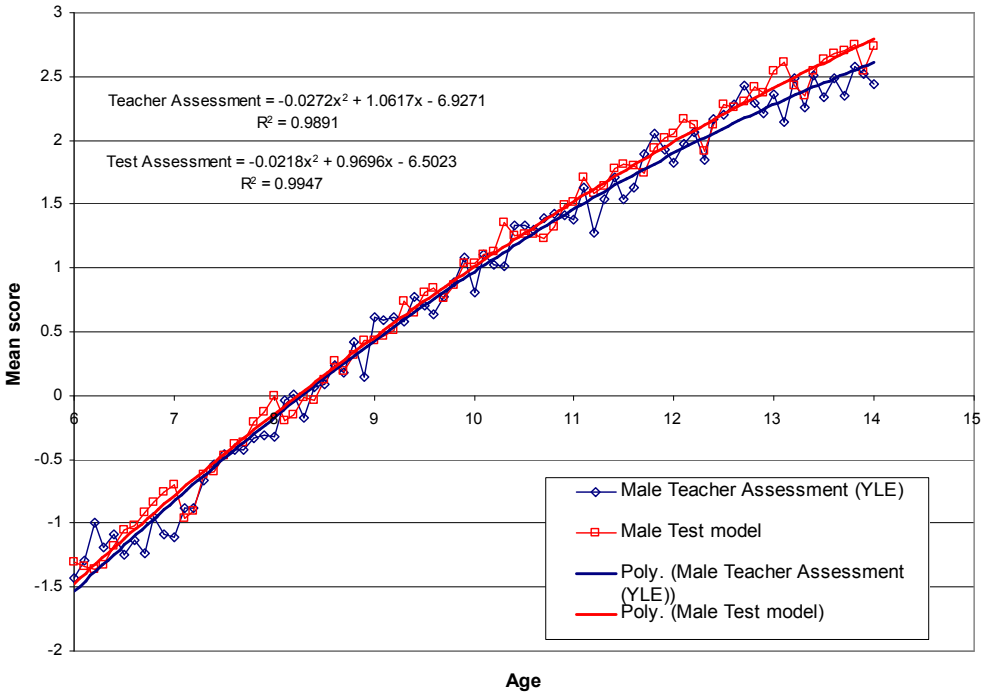


Figure 8.24 1998 Mathematics - Mean of Teacher assessments (Year level means equated) and Test Model assessments compared by age within Year level: Male students



Is the variability in assessment alignment a within-teacher or between-teacher effect?

The aggregated data show very strong similarities between test and teacher judgment assessments, particularly when trajectory differences are removed. However there is variation, illustrated early in the chapter and in the school site case studies, in the alignment of individual teachers to the test scale. From the case studies a high correlation of assessments from the two sources can apply even when invariance match rates are low. Even though only just above 50% of assessments are considered as matching, this implies a range of match rates for individual teachers, based on the assumption of a normal distribution of teachers around the mean rate. The data are very restricted (a maximum of 5 cases per teacher) and thus limited in the degree to which conclusions can be drawn. How feasible might it be to improve the alignment of teacher judgment assessments to the test scale?

Some broad speculations based on the 1997 and 1998 cross tabulations of match rates and teacher-test correlations for individual sites can be made. These are explored in Appendix 13. Only sites with more than 5 students assessed are included. This censoring reduces the bias towards high correlations when n is very small. Unfortunately it also eliminates small schools from the analysis biasing the analysis towards larger school sites. The match rates and correlations for the included sites are calculated and ascribed to the estimated number of teachers at the site and tabulated in Appendix 13. While assuming all teachers at a site are equal removes the between teacher differences at a site, it allows a rough estimate to be made of the proportions of teachers who have varying mixes of matching to the test and varying degrees of correlation to the test scale across the set of schools included. Over the full set of students with teacher and test assessments (1275 in 1997 and 2105 in 1998) an estimated 700 teachers are potentially included. Limiting the analysis to sites with more than 5 students reduces the number of teachers included to about 600. Summarised statistics from Appendix 13 are tabulated as shown in Table 8.8. Cut points for low/high are above 0.7 for correlation coefficients and above 0.4 for degree of match of assessments.

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

	Low Match	High Match	Total
Low Correlation	10%	30%	40%
High Correlation	10%	50%	60%
Total	20%	80%	100%

High correlation: 0.7 or above, Low correlation 0.6 or less. Coefficients rounded to one decimal.
High match: 0.5 or above, Low match 0.4 or below. Match rates rounded to one decimal.
N estimated to be 600 teachers in Years 3 and 5 for 1997 and 1998 combined.

Most site match rates are in the range 0.5 to 0.7 (see Figures 8.8 and 8.9). Only 20% of teachers are estimated to be 0.4 or below, in 0.1 match rate categories. Then, as a general estimate, about 20% of teachers have low match rates. About half of these are estimated to also have low correlation rates. The case studies presented earlier show it is possible to have moderate to high correlation coefficients and still have low match rates due to scale displacement. Accordingly, about 10% of teachers are estimated to have assessments with both low match rates and low correlation coefficients. The assessment approaches of this set of teachers would need to be better understood. They would be amongst the highest priority in developing strategies to improve teacher assessment calibration to the test scale. It might be established that a proportion of these teachers is unable to discriminate learning changes at all and thus cannot be calibrated to the test scale. On balance the speculative data offers optimistic possibilities for the improvement of teacher calibration to the test scales. Strategies that might achieve this include coaching, individual feedback on their current relationship to the test scale and specific training about the meaning and value of the scale in recording progress and intervention options. That consistency within a site can be established (even if miscalibrated) offers evidence for the potential to improve individual teacher calibration.

Summary

This chapter compared two independent methods of student assessment in two learning areas, to investigate the degree to which they appear to arrive at similar assessment results for individual students. The analysis of individual cases at Years 3 and 5 was expanded to compare samples of students assessed by teachers using a standard assessment framework compared to a model of test results for Years 1 to 8. In these comparisons the test and teacher-assessed 'samples' of students are notionally from the same Year level populations, independently sampled rather than being specific students with assessments from both sources.

Comparing teacher and test assessments for the two learning areas presumes both methods are assessing essentially the same skills and behaviours. The common patterns by gender and age within Year level within learning areas established in the analysis, offer support for the validity of the comparison. The clear differences between the results for English and mathematics across learning areas and the similarity of the results between methods within learning areas provide evidence that teachers, on average, produce aggregated summary level assessments that are measurably equivalent to the results from tests.

The links between the two assessments scales for each learning area that allow teacher and test assessments of student learning status to be placed on the same scale were established through a general transformation of the teacher assessments, using a Rasch model analysis.

This equating of scales is based on the assumption that measures of central tendency of the assessments of a large population of teachers are the best basis for equating the scales. Alternative methods of equating based on a series of specific behaviours and skills, a criterion rather than normative basis for equating, might establish a different equating relationship. Such a process was not feasible with this historic data. If such a criterion process were applied, the author assumes the result would affect only the placement of the control lines for the match zone, influencing which cases were deemed to match, not the general degree of match.

Once the assessments scales were equated, the two methods of assessment were compared for degree of match. On the basis of the estimates of measurement error for the two assessment processes, just over 50% of the assessments of students in each learning area are deemed as measurably invariant (i.e., within error) in Years 3 and 5.

The errors of measurement in both assessment processes reinforce that all assessments ere likely vary from the inferred 'actual' status by quite an amount. This illustrates the risk of using one-off assessments – whether by teacher or test. Regular re-assessment is one of the tasks teachers routinely do as part of their normal practice, except that a common vertical scale is rarely applied. Regular re-assessment by test processes on a vertical scale is also feasible, in the form of computer adaptive testing, but at a considerable additional cost.

There is evidence from the site case studies that a common assessment culture, to consistent criteria across a number of teachers, can apply within a school. This consistency applies even though the assessments themselves are not regarded as matching on the basis of the norm-developed translation of teacher assessments to the test scale. The general options of mis-match were speculated upon in Chapter 4. The case studies support that speculation. The quantitative process for articulating the differences between a teacher's set of assessments and the test scale provide a potential basis for improving the calibration of the teacher to the test scale.

Even though the student assessments may not meet the stringent criteria for matching, the correlation of the order of the students on the two assessments axes is often high. When approximate consistency of order applies the relationship of teacher to test assessments can be understood through the gradient and intercept of the comparison plots. Assuming a symmetric regression that allows error on both axes, the slope when near one indicates a consistent spacing of the assessments on both axes. Gradients markedly greater or less than one imply compression (or expansion) of the scale on one of the axes relative to the other. Zero or vertical gradients indicate no relationship of the two scales. Negative gradients imply a reverse order of students in the two processes.

The other aspect of difference, the intercept, indicates (based on its sign) that one assessment scheme is assessing above or below the other. A gradient close to 1 but an intercept markedly different from 0 implies a systematic displacement of the scales. This was illustrated in some of the cases studies and was consistent across multiple teachers at a site. Teachers in this situation are assessing the students with consistent order and spacing on the teacher scale as applies on the test scale but with systematically higher or lower values. Since the values on the scale have specific meaning in terms of what students can and cannot do, the displacement negates the interpretive value of the scale. The consistent order but displaced relationship to the test scale indicates a need to clarify the link between the scales. The consistency implies it would be feasible to recalibrate teachers to an interpretation of the level scale that is more closely aligned to the test scale. An implication of the consistent displacement is that fewer student assessments are seen as matched even though the teachers and the test are assessing consistently relative to each other.

The measurement characteristics of a test are fixed by its design. It works approximately the same in all applications. The teachers on the other hand are not automatically aligned to the test scale. Each alignment is a personal calibration. It would appear from the case studies that teachers could be aligned with each other and consistently to the test scale at varying degrees of displacement. A closer match would appear feasible through coaching, training and feedback from regular personal test comparisons for each teacher. Based on the estimates of individual teachers' correlations with the test scale, very few would not be able to be linked to the test scale. A universal alignment would seem possible for many teachers, subject to a prior step of re-examining and refining the scale structures.

The teacher assessments appear to show different perceptions of the trajectories of learning growth with age/Year level relative to the test even when the teacher scale unit is transformed to the test scale. This is an implication of the apparently differing trajectories of the test model compared with the transformed teacher data. The differences in the trajectories, if real, imply that teachers, relative to tests, underestimate skill levels for younger students and over-estimate skills for older students. More likely, however, is that the difference is a combination of inadequate modelling and a distortion of teachers' assessments due to the calibration issues discussed in the chapter. Resolving whether there is a systematic variation in the teachers' perceptions of the scales requires more teacher and test data at each Year level and a better study design. Such a study would be feasible in Victoria where data from both assessment processes are available at Years 3, 5, 7 and 9 at least.

Equating the trajectories for teacher and test assessments (i.e., eliminating the systematic differences) shows that the general features of the mean learning status as reported by teachers and estimated from tests are very similar. The gender analyses are consistent

between test and teacher assessments within learning areas. The mean learning status by age structures within a Year level are very similar with both showing a characteristic tail for over-age students. It is most unlikely that teachers consciously consider age and gender aspects when making their learning status judgement. There is a small over-estimation by teachers in favour of girls in higher Year level mathematics relative to test assessments.

In summary, an adequate view of population learning patterns by Year level and age and learning area can be obtained from systematic teacher assessments approximately standardised through a level structure. Whether a time series of an individual student's assessments can be used to monitor learning growth is less clear. Were it possible, the spin-off benefit of this to personalised management of student learning development would be high, through consistent and meaningful interpretations of what a given scale value implies about skills to be developed next. How teacher assessment for individual students might be integrated into an improved learning support and management system is addressed in the concluding chapter.

Chapter 9 Weaving the threads together

> Imagine an afternoon when a teacher can sit down at a computer desktop and quickly sort through reams of data she'll use to plan lessons for the next day… She'll compare every student's achievement against state standards to decide which students need review and which ones are ready to move on…That technological capability can only be found in the rare classroom today, but some experts say that such a data-rich approach to instruction will soon be common place.
>
> Hoff, 2006, p. 12.

Appraising the principal character

In Chapter 1 the principal character of the thesis, teacher judgement assessment was introduced. In the author's mind there was uncertainty about how the strengths and weaknesses of the character would play out. It was not then obvious whether the evidence would support either of the two propositions raised as the essence of the thesis, that teacher judgement assessment could provide valid indicators of learning consistent with test assessments. This evidence was to come from a range of sources. Part of the evidence was to be found in the early history of assessment. Part was to be found in previous and contemporary research on teacher judgement assessment and part from an analysis of unique data from South Australian teachers who used their on-balance judgments, in conjunction with the Statements and Profiles for Australian Schools' framework, to assess the learning status of samples of five students.

The main findings from the data analysis

Teacher judgements of student learning status in English and mathematics in the South Australian data have strong parallel relationships with test assessments of learning status in Literacy and Numeracy respectively. This applies in a situation where the teacher and test scales within each learning area were independently developed and applied, with no attempt to help teachers with the alignment of the scales through teacher training. Most teachers, however, were trained in the use of the teacher scale.

Year level means using the original teacher profile level scale have a linear trajectory with Year level. This is a direct result of successful implementation of the intended design for curriculum levels. The scale intervals at the design stage were descriptions of criteria developed in a strand over about two years. The consistent gradients with Year level that have been established are slightly less than 0.5 of a level per annum (0.472 to 0.468 in Figures 4.7, 4.8 for Victorian CSF/VELS Reading; 0.374 English; 0.41 Mathematics in Figures 7.3, 7.10 for SA profile levels) and are generally consistent over a range of Year levels. The teacher judgement assessment scheme was designed to work as a linear scale with time and

that was achieved very successfully. As such the mean values for each Year level can be seen as grade equivalents, the expected mean level scale value for each Year level. Year level means using the test scale, on the other hand, have a curved relationship with Year level. This raises both a complexity in the development of a scale common to both assessment processes and some fundamental issues about developmental scales. Once transformed to the test scale, teacher judgement assessment appears to document learning in essentially the same ways as do tests.

Based on modelled and actual test data for Year levels 1 to 8, the relationship of test assessments with teacher judgement assessments holds across 8 Year levels and thus across primary (elementary) and secondary teacher cultures. At a summary level the learning characteristics in English and mathematics by gender, age in 0.1 increments of a year and by Year level can be equally well described by each of the assessment processes.

Teacher judgement assessments, when transformed to the test scale, appear to follow trajectories of learning improvement with age/Year level that vary systematically from the test assessments. Teachers appear to underestimate the learning status of students lower on the scale and over estimate the status of students higher on the scale, in comparison to the test assessments. The lack of actual empirical data for test assessments at multiple Year levels leaves this as an open issue. When the apparent difference in trajectory is removed by equating the Year level means, the patterns by gender and age within Year level are consistent across assessment sources.

Assessments have maximum value for the management of learning at the individual student level. There are indications that holistic on-balance teacher judgement assessments for individual students match test assessments for just over half the students. That is, for students with both teacher and test assessments, only possible in Years 3 and 5, and applying a norm established translation for the teacher assessment scale to the test scale, just over half of the assessments match; i.e., are measurably invariant within measurement error. This establishes that test and teacher assessments differ by more than measurement error for just under half of the cases. But it does not indicate which assessment process is likely to be the better estimate of learning status.

The comparison however understates the relationship. The non-matching cases are not necessarily random or unordered when they are grouped into the sets of teachers within a school site. When the patterns of teacher test assessment relationships by school site are explored, teacher assessments for some of the sites correlate well with the test but are positioned on the teacher scale (converted to the test scale), such that they are consistently displaced above or below the norm expected relationship. At these sites however few of the

cases meet the criterion for a measurement match. This implies that the teacher order of the student assessments on the learning status scale is consistent, to varying degrees, with the test scores but displaced from them. This appears to be an issue of the relationship between the two scales for the teachers at some sites. Judgement assessments by teachers are ordered similarly to the test results but do not meet the measurement criterion for invariance because of the scale displacement.

Consistency of order but different scale values reflect the difficulty in ensuring that all teachers use the teacher assessment scales in the same way, that they arrive at the same learning status values for students at the same point in their learning. Part of the source of this variation is likely to be the lack of a calibration process where teachers could regularly compare their judgements with independent assessment results using a school or system wide common scale. There are indications that moderation processes within some schools led to a school wide consistent displacement from the test scale. This implies that at these sites teachers applied the level criteria consistently across teachers, within Years 3 and/or 5 within a school, confirming that a consistency of assessment had been developed. The displacement implies the need for a second step in consistency, that is, reference to independent assessments designed to help consistency across school sites. These independent assessments might be, but need not be tests.[28]

Overall the evidence suggests that many teachers can judge and report the learning status of their students using levels scales as accurately as can tests. The professional skill of teachers in doing this is under acknowledged. This skill has the potential for further enhancement and might lead to as good a documentation of student learning growth as do tests.

Based on the brief summary of the findings from the data analysis above, combined with the research review, it is possible to draw conclusions about the acceptance or otherwise of the two propositions from Chapter 1. The propositions are addressed generally here to set the scene for a more detailed commentary on the overall implications and possibilities, based on more comprehensive reviewing of the evidence. The bulk of the chapter considers both evidence and speculation as a consolidation of this research thesis.

---

[28] In Chapter 1 and Appendix 11 the normal criterion for the Rasch model (50:50) item success versus a most likely much stricter criterion applied by teachers is raised briefly. This needs further consideration in the ultimate alignment of teacher and test scales.

The propositions: findings

First proposition

The principal proposition was that teachers' judgements of students' learning status (scale values), in school systems where they have been applied, were valid indicators of student learning status for all students and for all teachers, and were already of such quality and reliability that classroom, school and system assessments can be based on teacher judgement alone.

The evidence from the historical development of assessment and the research into teacher judgement confirm, in general terms, that teacher judgement assessments can be a valid indicator of learning status. The evidence from South Australian teacher judgements, when treated as a source for understanding the general dynamics of mean learning status by age, gender or Year level, confirms that aggregated teacher judgements provide very similar understandings to those from tests.

However, applying a strict criterion of 100% of teachers being able to make valid learning status assessments in every case, the proposition is not accepted. There are some teachers, the percentage of whom it is impossible to estimate from the data, where assessments differ widely from the test assessments and where general displacement of the relationship of the teacher scale to the test scale, or a poor quality test assessment for some students, cannot be seen as the reasons for the difference. It is assumed that for a subset of these teachers at least, the teacher's on-balance judgement is not a valid estimate of learning status.

Second proposition

The second but weaker proposition was that teacher judgements had the potential to be enhanced to the point where their on-balance judgments of students' learning could be regarded as valid indicators of student learning status.

The evidence for the second proposition need not be as absolute as that for the first. Given the overall patterns in the data, it seems reasonable to accept that there is the potential to enhance the assessment ability of most teachers and thereby provide improved and valid estimates of student learning status from teacher judgement assessments.

Teachers within some sites, even though assessing consistently within that site (and thus having a high correlation with the test and by implication with each other), appear to generate assessment values that are so displaced from the normative scale transformation for the teacher scale to the test scale that very few of the assessments are measurably invariant across the two assessment processes. Their assessments show up as not matching, part of the approximately 40% of cases that do not match. However their assessment behaviour implies

that the potential for teachers with this assessment profile to be made consistent with an alternative scale range is very high. Teachers with assessments having high correlations to the test scale, even though no individual assessments may be regarded as matching, are likely to be the easiest to re-calibrate to a test scale. This is because they are already ordering their assessments consistently with the test assessments and disagree only on the actual scale values to be assigned for each student

Sufficient evidence exists to accept the second proposition in general terms. This leads to the issue of how the potential might be developed. In providing responses to the general questions posed in Chapter 1, this potential unfolds in terms of process redesign and assessment system changes.

Responses to questions posed in Chapter 1

What is the history of the assessment of students using processes that can be applied by observation and/or by comparison to described criteria (as distinct from pencil and paper tests)?

The use of teacher judgement, moderated by descriptive frameworks, performance criteria or examples with which a student product can be compared, has applied for at least 100 years. Chapters 2 and 3 indicate that in the very early stages in the development of standardised assessment practices, the process of assessment of developing skills (handwriting quality, general prose writing skills) was based on comparisons with exemplars and criteria. The exemplars were used as points along a scale of development and student examples were positioned on the scale, based on a teacher judgement of the match to one or more of the exemplars. For convenience of simplicity, efficiency and to allow statistical summaries, the assessment was recorded numerically as one of the scale values or an estimate placed between two adjacent values.

The spacing of the exemplars on the scales was carefully considered and spread using statistical techniques related to odds ratios and standard deviations, producing scales that had both order and relative spacings between exemplars. Some re-plotting of the data indicates that a reasonable approximation with logit scales can be established, indicating that the original criteria scales can be seen as having strong links to contemporary learning progressions also spaced on a logit basis. The general history of judgement assessments confirms that the concepts of how to scale learning development have been known for over a century. The most recent expressions (SPFAS) and their refinements of the last decade in Australia (VELS as one example) provide a basis for refining strand scales of learning development for better use by teachers in recording learning.

What does the research literature on teacher judgement as an assessment approach say about what teachers do and how well they do it?

The literature on teacher judgement assessments is rather meagre, especially relative to the literature on assessment generally and psychometrics in particular. The classroom assessment practices and the accompanying record keeping processes of teachers are infrequently documented. In proportion to the frequency of assessment events in classrooms, particularly teacher judgement assessments, the research base is small. Moreover, that research has some methodological difficulties.

In most comparisons of teacher judgement assessments to other independent measures of learning, there are the fundamental issues of response form and transformational problems to place assessments from different assessment process on to the same scale. A very small number of research cases avoid the transformation of scores by asking teachers to estimate student scores using the score framework for the test with which the teacher judgement is to be compared. While this avoids one problem it generates another. From the cases reported it appears that the teachers were not very familiar with the test or with the meaning of the scores. In some cases the teacher addressed the test as if they were the student whose score was being estimated. In doing this the teacher was not given advice about the relative order of difficulty of the items, resulting in a rather difficult task for the teacher. Even so the teacher estimates of the students' scores were close.

More importantly, it is unusual for individual teachers to be one of the units of analysis in a teacher judgement-test comparison. Accordingly the research tends to report, as does the analysis in this thesis, the aggregated or averaged assessments of teachers. An understanding of the proportion of teachers who assess in the same order as a test but who are displaced from it either through a parallel displacement of the scale or through compression or expansion of the teacher scale relative to the test scale cannot be estimated due to the constrained nature of the data. The proportion of teachers who may be naturally calibrated to the test scale or systematically displaced from it is not usually reported. Most often teachers provide only a small sample of cases and they do not appear to be part of a process of extended feedback of results over repeated iterations to see if calibration to the test scale can be improved.

There are also fundamentally different reasons for exploring the adequacy or otherwise of teacher judgement assessments. Researchers differ about why the quality of teacher judgement might be important. Research on the effects of teacher expectations upon student performance indicates that teacher expectations influence student performance (Jussim & Eccles, 1995; Rosenthal & Rubin, 1978). When these expectations are based on inaccurate assessments, particularly where learning status is underestimated, learning development is

depressed. On the other hand, inaccurate over-estimations appear to have the opposite effect, encouraging learning gain (Hinnant, O'Brien & Ghazarian, 2009). Teachers' misjudgements can have grave implications to the school success of the misjudged students, notwithstanding the positive effect for others.

From the author's perspective, improving the ability of teachers to assess accurately should at least diminish the negative effects of misjudgement. If an appropriate teacher development process were put in place, along with the refinement of meaningful learning maps, could teacher assessments be improved? In an ideal design this would be a two way process, with the differences between teacher and test assessments impacting both the test and teacher assessment processes. This does not appear to have been researched (or at least published). England has a wealth of data that in principle could be mined for the trend in the degree to which individual teachers might improve, or not, their match to Key Stage results over a succession of years. Whether the link to individual teachers is included in the data held by UK authorities has not been explored by the author but it might be one source for richer insights into the effect of feedback to teachers of student results, and whether the teachers agreed or otherwise with the test assessments.

The general impression from an inadequate research base is that some teachers are likely to match tests assessments well but that there is considerable variability in their holistic assessment skills. There is also a lack of assessment literacy, the ability to interpret assessment data (Stiggins, 2008).

What does analysis of the 1990s data from the South Australian adoption of national profiles (Curriculum Corporation, 1994a) reveal about the ability of teachers to estimate the position of students on scales described by increasingly complex learning behaviours??

The data summarised in Chapter 7 indicate that overall, teachers produce consistent patterns of regular linear growth in mean and median scale values for their assessments, as Year level increases. The trajectory for test means by Year level is curved with growth in mean score reducing with Year level. The judgements required teachers to estimate the last level achieved and progress towards meeting the criteria for the next level. The second data component, the progress within a level, was represented in the analysis as a decimal value to one decimal point. Progress within a level has been a controversial concept in level systems. Traditionally systems that use teacher judgement assessment have used a zone basis for representing the learning status, a rather gross unit of little value in a formative or informative assessment scheme.

In this study, teachers did not appreciate that their response represented a decimal value when responding. From their perspective it was just a representation of progress by clicking at a point along a line. The full spread of the progress line was used by teachers, including no

progress, with many of the 10 possible positions within a level approaching the expected 10% of cases at each point. As seen in Chapter 7 some response points were more used than others but all were used. Teachers appeared to be able to represent their estimates of learning status at a level of detail comparable to the level of detail provided by a test score.

The evidence from the teachers' responses is that, given a framework similar to a levels structure, it might be feasible to have teachers estimate learning status in about 0.1 logit increments (based on the logits of the test scale). While much more research into a scale at this degree of resolution is required, particularly the degree to which teachers can really discriminate learning differences, the principle that data could be established and recorded easily at this scale is confirmed. In practice this is a degree of refinement implying the ability to discriminate between learning status values about 5 to 8 weeks apart, in the Year levels 2 to 6, assuming 0.1 test logits represents $1/10^{th}$ of 2 years learning development.

What proportion of SA teachers were effective on-balance assessors of students?

This proportion proved difficult to estimate as the data for only a few teachers (those in very small schools) could be seen as separate data sets. The balance of the data could be explored at a school level as n responses from n/5 teachers as a group, within either Year 3 or Year 5. Based on the analyses in Chapters 7 and 8, just over 50% of student cases were regarded as matching. These cases would be spread over a much greater proportion of teachers, say up to 80% or so, with this group of teachers getting say 3 out of 5 assessments in the matching zone. However within the set of teachers where matching was low there were some whose assessments correlated very highly with the test estimates. In principle they were calibrated to the test scale but systematically displaced above or below the appropriate test scale position. Taking this set of teachers into account, an estimate of the teachers who were partially calibrated to the test (the measure of assessment effectiveness in this case) could be as high as 90%.

This estimate is of interest because it helps estimate the size of the task to have most teachers matching their on-balance assessment to a common scale. It seems feasible to improve the ability of this large set of teachers to make on-balance assessments. A deeper analysis and consideration would be required of the nature of the common scales, the appropriate units to use and the relationship of these scales to the vertical test scales. However, successful calibration training combined with ongoing reporting of test results in forms that could be used by teachers to compare their pre-testing estimates seems possible.

What do teacher-generated and test-generated data reveal about the learning development of students throughout their 12 or more years at school?

Observing the growth in learning of students, individually or as groups, over extended time scales is reported only rarely. Based on this study, the trajectories of level-scaled teacher judgement assessments and test assessments by Year level or age have different shapes. Both assessment processes might, however, provide a basis for a vertical scale for monitoring the learning development of students. The analysis in Chapter 8 indicates one basis on which the assessment process might be brought to a common scale. Whatever might develop as a future approach to resolving common scale issues, the concept of being able to record learning status in particular strands of learning on common vertical scales on either a teacher judgement or test assessment basis would enable detailed recording of student learning development. In principle such records, might confirm to all students, that they were increasing their stock of skills. As outlined in Chapter 1 this would depend upon the scale units being fine enough to indicate learning progress over the scale of several weeks.

The general trajectory of the mean of students by age or Year level provides one basis for estimating the expected growth at any point and the rate of this growth. A model informed by the rich individual patterns for all students, accumulated in a consistent way over a number of calendar years and linked where possible to teaching strategies applied, should provide the data for sophisticated analyses of individual patterns. These analyses are required to assist teachers with interpreting their monitoring of each student. The resultant models, using progress to date for each student drawing of a range of assessments, might then provide options and advice for teachers.

However the meagre public[29] data providing longitudinal records of student growth indicate that, on a test assessment basis, the trajectories of individual students follow quite different trajectories from the mean trajectory. The non-ergodic nature of individual trajectories was considered briefly in Chapter 5. The ECLS data (Tourangeau et al., 2006) indicate a wide range of trajectories between testing periods spaced from 6 months to 2 years apart. Some students show consistent incremental growth, some show sudden then flat growth, some show flat then sudden growth and all possibilities in between. Some of the variation is due, no doubt, to measurement error. The trajectories from the Suppes et al. computer aided

[29] Based on comments on websites for vertically scaled tests, suppliers' proprietary data are held but not released in the public domain. Some test suppliers sites, NWEA as an example, make their data available for further research. This would be one source for establishing the general variability in individual trajectories.

curriculum (Suppes et al., 1976), where progress data were taken in each computer session, show smoother growth with time, but very idiosyncratic patterns for individual students.

The path for each student is much more complex than just a mirror of the average. A deeper understanding of the dynamics of individual growth is needed to drive the knowledge base for teachers to allow them to fine tune their support strategies. More frequent status estimates for individual students are required to develop this understanding. Teacher judgement assessments are one potential source for these data.

Design elements for a teacher judgement assessment scheme

At this point in the consideration of an integrated teaching and learning system based on teacher judgement a brief summary of a draft concept is required. Further comments and responses to the remaining questions from Chapter 1 require a description of the concept of teacher judgment assessment developed from the evidence to date. The design acknowledges that teachers are the major participants in the education process. Assuming that the current paradigm of teachers responsible for students will persist, as against some alternative non-human computer based mediation of learning, teachers are the main agents for optimising the learning development of students. Principals, system administrators, testing companies, politicians and, in some cases, parents have little direct ability to support the learning of individual students. The professional role of teachers as managers of learning through monitoring individual student progress is made central. The design assumes that regularly recorded data on learning status would provide a basis for the better management of individual learning. A major source for that data would be the judgements of teachers referenced to scales of learning developed from IRT analyses of test items or tasks that reflect the increasing complexity of the skills being learned.

The concept that seems feasible includes:

The development of scales for strands of learning common to both test and teacher judgement assessments, calibrated with equal interval units, as the basis for estimating a scale position at any time.

Progress maps for learning within a strand. These maps would provide sequences of empirically developed skills[30] ordered and linked to zones on the scales to help teachers plan personalised instruction, make assessments and refine their estimates of the likely learning status range for a student judgement assessment.

---

[30] As defined in Chapter 1 'skills' is used generically to cover all nouns used to describe those elements that make up a description of learned attributes (skills, knowledge, behaviours, etc.).

Regular, simple, record keeping of all student assessments using teacher judgement assessed scale values for each student. The frequency of recording would be based on noticeable changes in skill level but would be expected to average out at about one new scale reading per student per strand on about a three weekly interval. The actual frequency would depend upon one of the issues developed briefly below, the highest possible resolution for detecting learning change.

Data recording and analysis systems for each teacher that are simple to use, with built in applications to analyse and present graphical patterns of development, drawing on empirical research and teacher enhanced knowledge systems. In essence the system provides patterns, diagnoses and suggestions for what next, based on the most recent trajectory for each student.

The concept of using assessment data to manage learning is far from unique. What is specific to this particular description of data driven management of learning is the predominant use of teacher judgement assessments recorded as judged test scale values (or a value convertible to the test scale). Under this scheme the frequent teacher judgement data points are integrated with any other assessments, including test assessments, using the common scale. The scale values encode what the student can do, and thus can be decoded by any scale user to describe what the student can do.

To achieve this general concept some of the matters to be resolved are indicated briefly.

Teacher test scale relationship-common scales?

While it is possible to convert teacher judgement assessments to test score equivalents (Chapter 8) and the reverse, the teacher judgment scales developed to date appear to indicate different trajectories with age/Year level than do IRT based test scales. The teacher and test scales do not have a simple linear relationship, as might a Celsius to Fahrenheit conversion. The essence of the difference is that current teacher judgment scales appear to illustrate the development of learning as a linear trajectory. Test scales based on IRT indicate diminishing learning growth with age/Year level.

The test scale is consistent with most vertical test scales based on item difficulty. Increments of growth with time diminish at higher Year/age levels. The patterns of mean assessments and their SDs conform to what is expected from mathematical modelling of the IRT trajectories (Chapter 5). The trajectory shows diminishing growth and reducing SDs with increasing Year level for an IRT difficulty scale. For teacher judgement assessments, the linear growth and increasing SDs with increasing Year level are consistent with a linear model. It is the unit intervals on each scale that determine the alternative trajectory and SD patterns, while recording essentially the same learning development.

This raises some design issues. Which scale concept should be favoured and what consequences follow? Clearly the CSF/VELS/SPFAS teacher scale works in practice. The relationship of the teacher judgement scale with an item difficulty scale is approximately linear for the mid section of the scale, but appears quite different for the upper and lower segments of the scale. This is not a new issue (Camilli, 1999; Camilli, Yamamoto & Wang, 1993; Hieronymus & Hoover, 1986; Petersen, Kolen, & Hoover, 1989; Schulz & Nicewander, 1997; Yen, 1986). The same phenomenon, learning growth over time, can be described in different units. The levels approach is shown already to work with populations of teachers. The design dilemma is which form of scale to choose. Are they both valid scales? If one is valid, the condition of equal intervals cannot apply on the other scale. Statistical summaries on one of the scales will be less valid.

A pragmatic solution would be to use the existing teacher scale designs because teacher judgement assessments will be the more frequent data points. Test or other standardised scores can in principle be converted to the teacher scale and given the lower frequency with which this will need to be done, it may be preferable to re-scale these lower frequency events. Translation of these scores to the teacher scale would be automated. An alternative is to design the teacher scales using a combination of Rasch scaled test items and Rasch scaled teacher judgements of the same specified items/skills. It is anticipated that some item/skill placement anomalies might arise as the combined scale is developed. In this design, scale unit intervals would be based on difficulty with the anomalies revealing the reasons for any test-scale teacher-scale differences. The scale design is left open but it is assumed a practical solution can be found. The solution has consequences to some of the other design issues.

A further issue is the stability of the teacher judgement vertical scale. Test constructs and item difficulties have been shown to be stable over extended periods of 20 years (Griifin & Callingham, 2006; Kingsbury, 2003). The stability of teacher judgements assessments is unknown although it is assumed that they would be continuously adjusted, by regular feedback, to remain linked to the particular vertical constructs developed.. How teachers' judgements change over the course of a career is currently unknown, as one example of a range of issue needing further investigation. It is assumed for early career teachers that their judgements would be refined over the first 5 years before they obtain some stability.

Progress maps as tools to planning, judgement and resolution.

The scales are the frames that hold the learning progressions, the progress maps, where skills are ordered and spread to match the empirically established difficulty to develop the skill. Any strand will be replete with these skills, many bunched together. An assumption in this design is that skills, like well behaved test items (Kingsbury, 2003), are likely to maintain

their inherent difficulty over time and place. The numeral learning data of Tymms (Chapter 5) indicates that these orders remain approximately constant across a range of cultures and offer some confirmation that learning progressions may in some cases be universal. This applies at least for some simple skill chains. The order of letter naming and letter recognition (Kerbow & Bryk, 2005) is essentially confirmed by Justice et al. (2006), suggesting a scale based on the difficulty order to learn letter names is also valid. While the examples are possibly weak foundations for a complete system, they both illustrate the wealth of existing data from testing records that could help create vertical scales as part of the knowledge base for teachers.

Progress maps help the teacher by setting the skill context in such a way that it is possible to assess, by an open set of processes consistent with the Rasch/Thurstone independence of instruments, a current learning status along the notionally uni-dimensional scale for the strand. The numeral assigned can be interpreted to say something about the student. The numeral is simple to record. In principle, in a well-prepared future world, the same numeral assigned by each teacher should have the same meaning. Versions of progress maps are already available in the Victorian system (Griffin, 1990; Forster & Masters, 1996; Rowe & Hill, 1996). The utility of progress maps, as supports for learning, is dependent upon the validity of the learning orders they document. Some examples of progress maps are developed by expert opinion (Popham, 2007). Given the general argument here that teachers are potentially, if not actually, good judges of learning development these maps should be good initial indicators of dependent skills. However empirical examination and confirmation of the orders of skill development based on difficulty (as by Bond & Bond, 2003) is required.

The unordered skill lists described for each level in current level structures make it difficult for teachers to estimate and record finer discriminations of progress. Recent improvements, such as the VELS progression points (Victorian Curriculum and Assessment Authority, 2006a), which divide the criteria within a level into subsets, still retain unordered lists within these subsets. If there *were* a likely order of expression that can be empirically established, it would help teachers in their monitoring and support of learning if this order were indicated. Item maps that explicate more subtle orders and skill difficulty relationships provide a sound basis for feeding back to teachers the likely order that students will develop the target skills in a learning area. For the purpose of this study, the point that learning progressions can be described is sufficient to establish the principle. Some problems of over-detail and information overload can be anticipated in refining the design.

Highest possible resolution for a learning scale

Assuming that the broader scale issues can be solved, a critical and related matter is the smallest perceivable change in learning status, the resolution of the scale. This is equivalent to deciding whether a ruler can be calibrated in millimetres or centimetres. If the smallest noticeable change in learning status is of the order of a change over two months, the notional resolution of current levels schemes scales using 0.1 of a level, the assessment process might not be refined enough to provide useful scale values for informative assessment.

As outlined above progress maps of skills, linked to specific (and narrow) segments on the strand scales, might help improve the discrimination of teacher observers. Discrimination of positive change in a 2 to 4 week period would be required. This implies a scale sensitivity of 0.03 to 0.06 test logits (based on the SA tests of 1997, 98), less than the currently estimated SEs for tests or teacher judgement assessments by a (very large) factor. It is unlikely that SEs can be reduced to achieve this precision. However is it possible that multiple opportunities to observe and engage with a student might reduce SE to some degree in the manner that increasing test length does, allowing then for higher precision estimates? A useful research question might be: How much improvement in a specific skill, say reading, is required from a given point before an experienced teacher can observe the improvement? If this can be established to be consistent across experienced teachers, and is found to be of the order of 4 weeks or less of learning, a scale with a smaller basic unit would seem feasible.

If the proposed teacher judgement assessment scales cannot be refined to this degree they would only be as useful for guiding the support of learning as are current tests. Both might be most applicable as summative assessments for extended segments of the curriculum rather than as short time-interval progress markers. Further investigation of teacher judgement assessment in the way proposed, for assisting with weekly decision-making, would not be justified.

Use of numerals to represent scale values

The question of whether numerals should be used to represent a position on a scale (and by implication a set of skills) is answered from the author's perspective by the utility of the numerals. This utility includes order, spacing, coded recording and the availability for statistical summaries, notwithstanding the varying (teacher versus test) unit issues raised earlier. Data in numerical form, assuming reliable and consistent development, have utility over time, teacher and location.

Other researchers do not share this view. Forster (2009), based on the work of Wiliam (1998), is concerned that using grades (these numerals would substitute grades in many contexts) as feedback on individual pieces of work may not focus the student on what needs

to be improved. Butler (1988) reports that marks or grades engage the ego and can distract students from other supportive and constructive feedback. The Butler research was based on versions of conventional grading, in the social and personal classroom context that these create. It is not clear that the same dynamics would apply in a new context. In defence of the position proposed here, the scale value has meaning (more so than conventional grades or marks) and builds on engaging students with the criteria so that they are aware of the skills being developed, the standards required and as a prompt for self-assessment. Whether the potential negatives outweigh the positives could only be resolved by further development.

The use of numerical values to locate students on developmental scales is less of an issue in test assessments and is a given for most test schemes conducted to estimate the learning status of individual students. The test scale values have the general utilities of order, spacing and thus applicability to statistical summaries. A position along the scale for a given student (within an error range) is a major product of the test analysis process. The value of the position identifies where, in the myriad of skills to be developed, the student currently sits and is based on the students' responses to difficulty ordered items. With this knowledge a teacher has the information to focus on Vygotsky's concept of the Zone of Proximal Development (ZPD) for the student to optimise learning (Rogoff, 1990). A similar process, based on numerical values, should be able to apply for teacher judgement assessments.

What numeral structure to apply

A range of numerical conventions can be applied to the design of the scales. The VELS/SPFAS level scale assigns a zero origin and develops the main scale in integer increments. Individual levels can be subdivided into zones or fractions. The main teacher judgement assessment scales in operation (the VELS scales) currently use 0.25 of a scale level increment, in contrast to the original three zones in the CSF. One specific application within VELS, the English Online Interview (Department of Education and Early Childhood Development, Victoria, 2009a) uses 0.1 increments for one report to teachers. The SA data in Chapter 7 were recorded at 0.1 level increments.

Test scales tend to use positive numerals (transformed from logits) but with less intuitively useful values than levels schemes. The test scales have less direct meaning to teachers, initially at least, but with regular use test scale values would acquire meaning. A new language would evolve quickly; instead of a skill being about level 1.1, it would be, say, a 305 skill. The selection of the best structure, one that teachers would respond to intuitively, is a key issue but not one addressed in this concept description.

The data required to do this would come from the integrated observations made by teachers. On the evidence of teacher judgements in 1997 and 1998 it should be feasible to develop processes that increase the consistency of teacher judgements across classes and schools, to observe and articulate the learning development of students in a set of strands of English and mathematics learning at least (or in whatever re-structuring and re-labelling of these key learning areas applies from time to time).

If the scale and the ordering of the skills in developmental order is accepted by teachers, the position of a given student can be estimated in relation to the general spectrum of acquired/developing skills on the scale. It is assumed that teachers hold implicit hypotheses on the learning status of all students on a daily basis, even in the absence of a scale to articulate efficiently those hypotheses. A language is needed to express or communicate the hypothesis. The simplest form of this language is the scale value (or scale region) represented by a numeral. The hypothesis can be recorded from time to time as a data point.

When a teacher decides to record a data point for a student, the teacher judges what skills the student exhibits and can then place the student at the appropriate point (or zone) on the scale and record this value (or the midpoint of the zone) as the scale value. In principle the estimation process is efficient for the expertly trained and the recording process easily made on the fly as required, without requiring teachers to redirect teaching and classroom time to recording. Using a shorthand notation should reduce the recording time for expert and confident teachers relative to detailed checklists of skills developed to date. This general concept for a learning and assessment system based on teacher judgement assessments sets the scene for completing the consideration of the remaining questions from Chapter 1.

Addressing the remaining questions from Chapter 1

Assuming some teachers are relatively effective on-balance assessors, what tools and processes might be required to maintain and enhance their skills and develop those of less effective assessors?

The evidence suggests that a reasonable proportion of teachers are either effective on-balance assessors, or could be calibrated to be so rather readily given the development of some required tools and support processes. Progress maps aligned with developmental scales could be used to support learning status recording.

Once appropriate scales were in place, continuing with standardised assessment processes that could be used to independently establish the learning status of each student would be desirable. These would need to be available for regular use, be computer administered so that the results were available immediately and be reported using the common scale. Teachers

would be encouraged to make on-balance assessments and compare their assessments to the test assessments. The New Zealand asTTle process (Hattie & Brown, 2003), where teachers use their professional skills to specify test parameters and content, might serve as a model for doing this. In this model the test specification is a further expression of teacher judgement, through the requirement to target class needs in the specification, providing an additional feedback loop to teachers about their judgements.

NAPLAN scales, if national testing continues, could be brought to (or be convertible to) the same common scale as proposed for teachers allowing all data about an individual student to be recorded in a time-stamped common form. In its simplest form the interaction of adaptive testing, national tests, teacher judgements and within school and within district moderation processes should provide the critical mass of triangulated assessments to begin to bring each teacher's assessment to a common calibration range.

How might the design of classroom and school processes be changed to optimise the use of teacher judgements?

There is pressure on teachers to use data to better manage learning. US teachers, assumed to be indicative of a broader than US issue, allege that they "were not taught how to use data to differentiate and improve instruction and boost student learning" (Duncan, 2009, p. 3). It is likely that anywhere this pressure is perceived to apply a similar concern will be expressed. The use of data is a complex issue and as was argued in Chapter 1, traditional grades and marks as one possible source, do not provide adequate data to monitor student learning. Detailed checklists of skills achieved while indicative of how learning is developing, can be cumbersome. Research has not served teachers well to date in collaborating with them to develop simple, practical, sound processes to assess students and then to record these assessment in a form that they can use as data. Pressure to analyse current grades, marks or checklists better will not provide a complete basis for meeting the ideal of improving instruction and boosting student learning. A process that develops adequate longitudinal data to monitor learning status is required. Teachers should not be blamed for their current lack of skills in using data, nor should the institutions that trained them, when the concept of the appropriate data is still ambiguous and unresolved.

Based on the literature review and the data analysis, improving and standardising teacher judgement assessments may be a process that provides the required data. Particularly if through the development of common scales for expressing the value of the assessment, all forms of assessment including tests can be integrated into the one data system. Criticizing teachers, and they criticising themselves, is unjustified until a practical data system for teachers has been developed. Alternative scoring initiatives (Marzano, 2000a) using rubrics and improved data concepts, or alternative assessment planning processes (Biggs & Collis,

1982) only partly address the issue. To borrow from Fullan et al. (2006) a breakthrough is required. One element of the breakthrough is the acknowledged ability of teachers to know their students. However, to achieve a good understanding of students' individual learning, the total student load (Ouchi, 2009) needs to be below 80[31]. Teachers can make judgements only where they have adequate opportunity to observe students and develop individual relationships.

Teacher judgement would appear to provide a legitimate basis for developing an understanding of student learning development and for creating data points to monitor student development. Data for each student in the form of data points over short time intervals is recognised as one of the mechanisms to help teachers improve the targeting of their educational management of individual students (Timperley, 2009; Fullan et al., 2006). There is a tendency for commentators to see these data points as requiring an assessment process that is external to the classroom or school. There is a strong impression in much of the assessment literature that real data points require tests and only tests. Teacher on-balance judgements however appear to provide a basis for generating many of the data points. An integrated assessment arrangement would allow data from multiple sources to be brought to account. The much less frequent test assessments, as might be available, would help maintain teacher calibration and help to improve the learning status estimations through triangulation.

An understanding of the ways teachers' observations can be made part of the general classroom culture has not been well developed in the research literature. The concept of aligning teachers judgements to the scales used in tests is considered rarely. (Even though VELS was an example in operation, this aspect is now compromised by the adoption of the national scales). Accordingly, there appears to be little work underway on how teachers can be used as the main source of developmental data about individual students using scales common to tests and teacher assessment. Were the issue to be explored and found to generate reliable learning status estimates, it could diminish the priority given to tests and reduce long-term dependency on test assessments as the only valid measures of learning. In principle, an implication of integrating estimates of learning status through holistic on-balance teacher judgement is that a wide variety of processes ought be able to estimate learning status, just as a wide variety of rulers, as well as perceptual judgement can be used to estimate height or distance.

---

[31] For primary teachers this already applies with total student loads (TSLs) usually below 40. For secondary teachers the possibilities for teacher judgement assessments as data for individual student trajectories are reduced as the TSL exceeds 80.

The closest independently developed concept so far discovered by the author that is similar to the general outline above is the general model for data informed instruction described by Fullan et al. (2006) as part of their description of what is required for a breakthrough in learning management. In that outline they describe four core ideas based on, among other elements, CLIPs (Critical Learning Instruction Pathways) their terminology for learning maps.

The four core ideas are[32]:

> 1. A set of powerful and aligned assessment tools tied to the learning objectives of each lesson that give the teacher access to accurate and comprehensive information on the progress of each student on a daily basis and that can be administered without unduly interrupting normal classroom routines
>
> 2. A method to allow the formative assessment data to be captured in a way that is not time-consuming, to analyze the data automatically, to convert it into information that is powerful enough to drive instructional decisions not sometime in the future, but tomorrow
>
> 3. A means of using the assessment information on each student to design and implement personalized instruction; assessment for learning is a strategy for improving instruction in precise ways
>
> 4. A built-in means of monitoring and managing learning, of testing what works, and of systematically improving the effectiveness of classroom instruction so that it more precisely responds to the learning needs of each student in the class (Fullan, Hill & Crevola, 2006, p. 80)

Idea 1 is met in the teacher judgement concept. Teachers make their judgement based on a set of observations, assisted by a range of potential assessment tools and strategies. At any point where they need to crystallise their views for given students they consolidate their assessment into scale estimates. While they might consider the issue on a daily basis (Have I noticed something new?) they would probably crystallise their judgement into a data point only every week or so for any particular student, even though for convenience and efficiency they might do this for say, three to five students a day (as part of spreading the load for their focus and for their record keeping). Fullan et al. encourage simple daily assessments. The judgement assessment is expected in most cases to be done in such a way, and using such tools, as to avoid "unduly interrupting normal classroom routines" (p. 80.). The teacher judgement concept meets the first idea of Fullan et al..

---

[32] Punctuated as in the original, no full stops after each idea.

Idea 2 requires a non time-consuming process for data capture. The teacher judgement concept achieves this through the estimate using numerals. The score is easily entered by a calibrated teacher on the fly into a database assuming, for example, a wireless-based hand-held tool. Based on pre-designed models for analysis and charting, individual students can be reviewed and, based on their current status and time since last noticed change (all automated), advice about specific instructional strategies for the current scale location offered. In cases where new test or massed data become available, the data could be automatically added for each student, at the time/date of the test on the same scale as the teacher is using, and seamlessly included in the analysis process. In addition, classes where computer adaptive/targeted testing is included, the data management system would automatically update student assessment records. This would flag cases where teacher and test disagree, for re-consideration without any need for active data entry by the teacher. Where estimates agree this would be a consolidating event for both teacher and student. Idea 2 is met.

Idea 3 requires a view of the data that is personal to the trajectory of each student. This is addressed in idea 2 as an assumed efficient next step as new data are added or the learning status reviewed. The teacher judgment concept implies a personal focus on the data history for each student as part of a decision system of what to do next with each student. The expectation is that data are fed through to an expert system where data histories of large numbers of students are recorded. These data would feed to mathematical models that would report graphically to the teacher and offer prompts to teachers on what instructional experiences might be useful if progress did not seem be occurring. This might also moderate undue concern and anxiety about slow progress at some points. The knowledge base would highlight known consolidation stages.

The final idea, idea 4, requires the teacher to report what the outcome of any specific instructional strategy was, as part of the expert system for improving the effectiveness of the suite of strategies. Independent of any specific advice or comment added, the next reported learning status itself would indicate whether the teacher assessed the instructional strategy as working. A positive change in status, when time between points is considered, is an indicator of the possible impact of an instructional strategy, assuming the teachers had reported to the knowledge base what they were intending to do next.

Taken as a set, all the requirements of the Fullan et al. breakthrough outline are met by the teacher judgement assessment concept. Data and objectives are connected. Learning data are observed and recorded efficiently and, as required by the teacher, not as an external pressure.

Of course it is likely that teachers might be required to comply with some internal school schedules for assessment data[33].

The recording of a learning status estimate offers the possibility of immediately suggesting instructional strategies and refined assessment possibilities back to the teacher, to assist in the management of each student's learning. Thus the teacher knows at any time the approximate learning status of each of the students in a coded form that has meaning for both the teacher and student (once the scheme has been running). The professional judgement of the teacher is enhanced, refined, calibrated and supported by the data analysis systems, the knowledge base and the expert advice system. The teacher owns the data and has a keen interest in crosschecking, confirming and updating as each interaction with the expert systems offers confirmation and options to consider. The major source of the data is the teacher whose self-esteem one assumes will be enhanced when teacher judgement and tools provide similar perspectives. Where they do not the teacher is prompted to double-check.

A further spin off of the data management process is the potential for automated procedures for drafting summative reports to parents. Such a process should reduce (but not eliminate) the time required of teachers in report production. The data would also be in a form that would allow on-line parent access to data about their children. Whether this is desirable is an independent question but such access already applies for some school systems for grades and assignments. This concept changes the nature of the data reported and adds the potential for meaningful progress reports. It also adds a strong feedback driven incentive to standardise the scales across teachers within a school.

There are also implications to the way in which the operation of a class or Year level is managed. Small groups would be an efficient process to support targeting instruction to sets of students of roughly equivalent learning development status. As indicated by Fullan et al. (2006) and earlier by Fitz-Gibbon (1992), the use of peer or cross age tutoring or students working in pairs might be other strategies considered to make the intention of the teacher for personalised development support a practical possibility. An initial promotion of practical classroom strategies with no greater demand on teachers than currently apply would be

---

[33] See a sample whole year reporting schedule and extracts from whole year assessment calendars developed by Hampton Primary School (Hampton Primary School, 2006) as an example of a highly structured internal assessment schedule. While a school wide process is required a teacher judgement scheme with data sent to a common database in a common format might reduce the number of required lock-step common assessments. Some other aspects, such as reporting to parents, might be also be simplified.

required. The proposed knowledge base, as it evolved, would provide teacher developed support that would effectively self-manage the options for class processes.

Current implementation of elements of teacher judgement assessment consistent with the thesis

Some elements of the suggested ways in which teacher judgement assessments could be used are already in place or are under trial. Teacher judgement schemes apply in England, Scotland, Wales, New Zealand and Victoria as four examples. Based on information from the Qualifications and Curriculum Development Agency (QCDA) and National Strategies websites (Department for Children, Schools and Families, 2009; Qualifications and Curriculum Development Agency, 2010), England has been trialling a new approach to assessing the progress of students, described as assessing pupils' progress (APP). Classroom assessments of learning status in the trial primary schools are holistic teacher judgements.

A "sub level", the approach in England to progress detail within a level, is assigned by refining a judgement through reference to criteria just above and below an initial judgement. The intention is that teachers use the assessment process to fine-tune their understanding of learners' needs and then tailor their planning and teaching accordingly. Diagnostic information about students' strengths and weaknesses is used to modify teaching and improve learning. As a result, teachers are expected to make reliable judgements related to the national standards by drawing on a wide range of evidence leading to assessment data that track pupils' progress.

Based on the trials of the teacher judgement process (Qualifications and Curriculum Authority, 2009a) the feedback from the evaluation (of Assessing Pupils' Progress-APP) was positive and supportive of holistic teacher judgements.

> Most teachers considered that the use of APP had improved their ability to identify gaps in pupils' learning and also reported that they found it easy to make the link to their planning so that APP assessment outcomes could inform next steps in teaching and learning. There were positive comments about how APP complemented the new frameworks. They also felt that they were better able to identify 'naturally occurring' assessment opportunities and their questionnaire responses showed a growing trend in the use of observational assessment. This was welcomed by many as an opportunity to improve classroom practice in year 1, building on the strengths in assessment from the early years foundation stage (EYFS).
>
> A number of teachers and headteachers reported that they were intending to replace at least some of their existing assessments with APP, as this would give them a more accurate and holistic picture of pupil attainment.
>
> Headteachers and local authority staff emphasised the improvement in teachers' confidence in their own ability to make accurate assessments without the need to rely on a test or assessment task and said that teachers felt empowered by this.

> Local authorities were clear that the use of APP promoted more sharing of responsibility for attainment and progress across key stage 1. (Qualifications and Curriculum Authority, 2009a, p. 8)

The approach adopted in England has some of the elements considered by the author earlier in the chapter. The assessments use a lower resolution assessment scale than the author posits is feasible. Whether a more refined scale (below 0.1 of a level) would work could be established only by further trial development. However the trial suggests that teachers are finding the general model both attractive and useful.

The Victoria, Australia school system has developed a levels approach combined with teacher judgement assessments. The division of the Victorian Essential Learning Standards as discussed in Chapter 4 and earlier in this chapter, into decimal progress stages (0, 0.25, 0.5 and 0.75), confirms a move away from descriptive zones within a level to a numerical representation of the progress. The progression points as numerals provide evidence that some elements of the model proposed by the author have already been developed (Victorian Curriculum and Assessment Authority, 2006a; 2006b). Student learning status is recorded as a level and at a point of progress to the next level, in a numerical form.

More recent developments include the releasing of conversion scales so that NAPLAN scale scores can be converted to VELS equivalents. This allows schools to maintain a commonly structured scale, starting at below 1 (possibly 0) and extending above 6. The VELS equivalents of the NAPLAN scale are not regarded as exactly matching the previous VELS scale up to 2007 (Victorian Curriculum and Assessment Authority, 2009). Using the VELS equivalence scale, schools appear to be converting their NAPLAN scale data to the VELS scale to allow a better scale format with which they can summarise Year levels between those tested and have an approximate link with previous data summarised in VELS units. This observation is based on a small number of examples with web published Annual Reports (Caroline Springs College, 2008; Marist-Sion College, 2008). The conversion scale appears to be released privately to schools, thus it is difficult to establish public domain detail of the conversion. The process of maintaining the common scale adds confirmation, if it was needed, that a consistent scale over Year levels and over calendar Years has value to schools.

The range of tools appears to be developing for English and mathematics (Department of Education and Early Childhood Development, Victoria, 2009a, 2009b) and the link to the VELS scales seem to be maintained in the face of environmental changes such as the introduction of the NAPLAN tests on a different scale. While the England and Victorian systems are incomplete expressions of the combination of scales, teacher judgement assessments using scale values, progress maps, test data recording and data analysis, they include many of the elements that begin to meet the integration of test and teacher

assessments outlined by the author. There are indications that the thought experiment results have some applicability in the real world.

What options might need to be considered for those teachers who have limited abilities in on balance judgement?

In mind in framing this question was the issue of the teacher, who after an opportunity to attempt to develop their judgement skills, drawing on what ever resources are developed to assist this, could not estimate the learning status of some known cases (in vivo or through video examples). This also assumes that most other teachers exposed to the same assessment development options had improved their judgement skills. A teacher who continues to be unable to make a reasonable estimate of learning status (in the context of most other teachers now being shown to have improved their ability to do so) might be considered as also unlikely to know what to do to assist students. This assumes teachers can assist students effectively only if they can establish their students' current learning status.

Given the assumptions about how a teacher judgement system might work with such regular case-by-case feedback, it is difficult to imagine teachers who could not improve their assessment skills. However for the teacher unable to meet certain criteria for assessment after an acceptable period, and with specialised support, it should be clear that this is a teacher who is unable to personalise support to students and who has an inaccurate view of current learning status and student needs. One assumes that at this point the teacher ought to be counselled to seek other employment, or contribute to education in some other way.

If some forms of teacher performance criteria are to be developed, a better basis than just the mean learning status improvement of classes or schools is required. Teachers' abilities to estimate learning status might be a better basis. Building criteria for effective teachers based on this ability, with supporting tools and data systems, may be more acceptable to teachers. The effective teacher would be one with good assessment skills and good instructional strategy choices, leading to a regular rate of learning improvement. The rate of improvement, the skill in assessment and the background characteristics of the students could be bundled into a more comprehensive and total quality learning management system. This might be a more productive approach to teacher quality and performance than basing the assessment of teachers on test results, or test results alone.

What would be the implications of the proposed designs to teacher pre-service training?

As far as can be established there is only limited training or preparation of teachers in student assessment generally, what Stiggins (2008) terms assessment literacy. Stiggins argues that

> Such literacy is needed to design and build totally integrated assessment systems with all parts working together in the service of student success. While virtually all [US]

> state licensing standards require competence in assessment, typically neither pre-service nor in-service teacher or administrator training programs include this kind of training (Crooks, 1988; Black and Wiliam, 1998; Stiggins, 1999; Shepard, et al., 2005). (Stiggins, 2008, p. 11)

It also is most unlikely that there is much teacher preparation in making on-balance judgements as part of the assessment training for beginning teachers anywhere in the world. It would require a separate thesis to establish what is included in any current assessment training for new teachers. The UK Professional Standards (Training and Development Agency for Schools, 2008) and Victorian Institute of Teaching standards for graduating teachers (Victorian Institute of Teaching, 2009) provide an understanding of the intention for initial teacher skills in these two educational jurisdictions. These are appropriate sources since these two regions appear among the front-runner implementers of teacher judgements as one element of an integrated assessment system, as described earlier in the chapter and in Chapter 4. In both cases there are strong emphases on effective assessment processes and knowledge as part of the standards to be met by new graduate teachers. Neither standard specifically mentions teacher judgement assessment although given the broad nature of the standards it may be reasonable that they are described generally without specific detail of appropriate assessment approaches. If the arguments of the thesis were to be carried forward, or even to ensure that the teachers are appropriately trained, standards of this sort would need to be more explicit about developing teacher judgement assessment.

It is not surprising that teacher judgement assessment does not yet appear to be a well-described process for assessment in teacher preparation or as part of professional standards. The evidence presented here suggests, however, that it has a strong contribution to make to the development of an integrated assessment system that optimises the value of the professional skills of teachers. It interacts with, and has the potential to provide the data for, longitudinal tracking of student development across strands of subjects. The development of the skill of interpreting student assessment data, particularly through the use of longitudinal models and data mining processes custom-built for schools, is a requirement for emphasis in future teacher education standards. This is consistent with the Duncan (Duncan, 2009) assertion raised earlier that US graduate teachers complain that how they should use assessment and other data is a missing component of teacher preparation. This thesis argues that both the creation of the data and the interpretation of the data about learning need emphasis, to improve the quality of the learning experience for students.

All this presumes the development of an integrated system of clear, agreed curricula for schools with a developmental description not premised on a lock-step view of all students achieving predesignated outcomes at specific Year levels. Although the achievement of all

outcomes for all students at the same time would be desirable, the evidence based on age patterns and idiosyncratic individual rates of learning suggest strongly that teachers and schools cannot achieve this. To keep track of student development and to help teachers optimise the learning of individual students, a framework describing learning outcomes in the form of key milestones in strands of the curriculum, freed from the Year level structure is required. A vision for learning and the assessment of learning that emphasises making the growth in learning visible to the learner, the teacher, the caregivers, the school and the system is also required. Based on the potential skill of teachers being able to judge students' learning status and their being able to integrate information about each student from multiple sources, this integrated system should be built on the teacher as the centre of the assessment process.

Support and tools necessary to fulfil the teacher judgement assessment process with much greater refinement than currently occurs will be required. As described throughout this thesis the teacher is already at the centre of learning management for students. The required vision, tools and training for the role to be carried out better, are missing.

One strategy for teacher training would be to include the observation of a small set of students over the period of training, to build observational and information integration skills. A focus on a set of say five students observed over four years should set the style for these new teachers as they recognise the new student skills developing at each observation. The further implication is that through this process, if the future teacher demonstrates an inability to understand and articulate the learning status as the observation periods continue, grave questions about the value of her/his continuing as a teacher should be raised.

Assuming both the appropriate breakthrough evolution of teaching and learning as outlined by Fullan et al. (2006) and the integration into this process of the observations of students by teachers as the prime evidence of learning, the training of teachers will need its own breakthrough to develop teachers compatible with the new skill profile of teachers. A requirement to demonstrate the required skills will modify the development and selection of teachers. The teachers so developed should make a difference to every student's life, partly because they will understand where each student is in the educational journey.

In conclusion - the fate of the principal character

Teacher judgement assessment has great potential. It has been shown throughout the thesis that teachers, given appropriate frameworks, encouragement and support tools can integrate their observations and other data to make estimates of learning status that parallel learning status estimates made through test processes. For some teachers their estimates match the test estimates closely. For others the order of students is approximately consistent but the test and the teacher scale are widely displaced. These teachers have great potential to be able to align

with a general model of learning development in an appropriate scale relationship. It is only their internal understanding of the scales that are slightly awry.

A small set of teachers may remain who do not have the understanding of learning nor the observational skills to estimate the learning status of students adequately and consistently and do not improve these abilities with regular feedback and training. The truth of this can be established only by appropriate research and only in systems where the general model of teacher judgement assessment with independent support and feedback is already being developed. The appropriate treatment of this small set of teachers, if their existence were to be confirmed, might be to remove them from the classroom in the interest of student learning until all teacher development options have been exhausted. If, after these options are exhausted, the teacher were still unreliable as an assessor, return to the classroom would be unfortunate for students. They will be less able to make appropriate decisions about how to support and manage the learning of most or all of their students. A sound education system would want to avoid this situation.

A system built upon the professional judgement of teachers and with support arrangements to keep this judgement tuned, will provide many benefits. These benefits will be to students in other ways than just a teacher's understanding of them. They will be to parents and care-givers, to other teachers and through summarised information, to schools, districts and systems in the ability to better understand what is happening in the learning development at each level and for each audience.

Currently education systems run the high risk of further de-skilling and demoralising teachers by allowing regular and consistent messages about the lack of professional judgement skill in teachers to go unchallenged. The insistence on tests as the only process to ensure the quality of teachers, or as it is sometimes cast, the only way to know what is really happening to children, is clearly inappropriate. Tests have their value and that value is to teachers themselves, as one source of feedback on their judgements. But tests are unable to manage sensitively the learning development of children.

The feasibility of a teacher judgement assessment approach to managing student learning is dependent partly upon further research on the current judgement skills of teachers and how these judgements might be enhanced. One avenue for immediate research would be the Victorian school system where teachers' estimates of learning status on the various VELS scales, just prior to NAPLAN tests, could be compared with NAPLAN results. This investigation would have the advantage of an already understood common scale. More generally, teachers in other Australian school systems could be invited to estimate student scores using the NAPLAN scale and these compared with actual results. Training over two or

three years could be explored to observe whether the match of test and teacher estimates improves.

That it should be possible to educate teachers to better estimate the latent attributes of learning in a number of strands is supported by the evidence and cases presented. The analysis of Hubbard (2007) of the benefits of training in the measuring of intangibles supports training as a process to achieve this. The essence of the feasibility of estimating learning is whether it is possible to detect differences between different successive states of an individual student's ability to do, say, think or perform a task as a subset of the possible changes from time 1 to time 2. If differences can be detected, is it possible to scale them? Being able to observe the difference may for many observers require experience to develop the connoisseurship skills and tacit knowledge and ultimately a common calibration. However, if most teachers can detect the differences, the basis is there for building many elements of the common scaled teacher judgement assessment /measurement system.

The interaction of teacher judgement possibilities with the test development world is another important consideration. Considerable investment (and profit) applies in the provision of tests and related data systems. The public good would be most served where all such systems were required to indicate the scale relationships of their products to the scales that would eventually evolve for teacher judgement assessments. The logic of this is illustrated by the Victorian approach to re-scaling NAPLAN scales to the VELS scale to maintain links to previous data and to maintain a common approach across Year levels. Given the reach being developed by some test publishers through acquisitions and the threat to the volume of their business if processes were to develop that diminished (the possibly hoped for) dependency of teachers on external tests, strong technical arguments against the validity and reliability of teacher judgement are likely to arise from this interest group in particular.

In a review of the early years' assessment of English in Victoria, Care, Griffin, Thomas and Pavlovic (2007) consider, among other matters, the inability of one test a year to provide the understanding of how a student is developing.

> It is the view of many researchers and policy makers (eg. Paris & Hoffman, 2004), that a single assessment cannot represent the complexity of a child's reading development. The most valuable assessment will provide evidence about a child's developing skills that will demonstrate growing competence, as well as lend itself to comparison against normative standards of achievement. This is an approach that is used intuitively by classroom teachers, who collect and integrate information across a range of reading factors. It is a useful model to consider. (Care, Griffin, Thomas & Pavlovic, 2007, p. 45)

Useful indeed. But in addition to integrating the information is the need for a more universal language for sharing the information within each learning area. Many local languages exist;

in a class, within a school, within a set of like-minded teachers, within some school systems. One approach to finding a more universal but concise language might be to refine it from the developing understanding of scales for constructs, generalised from psychometric research. Levels are a beginning. They have provided the basic 'tick marks' of the scales, but some filling of the spaces between them is required.

The general concepts described have developed as a result of the exploration of the history of teacher judgement assessment approaches and evidence so far. Nowhere in the literature is any author rash enough to propose such a general model along the lines that are developed here. This is, one assumes, not because many others have not thought of it but because they have had the wisdom to find the flaws so obvious and overwhelming as to not to bother to develop their thoughts further. However in this thesis teacher judgement assessment has played its role and is found "to pull its weight". No competing characters have the potential to develop their purpose as effectively or as economically. Not only does it justify its role but with careful encouragement, it offers a scaffold for system learning as well as personalised student learning, one key element of the much-needed breakthrough.

References

ACT Board of Senior Secondary Studies. (2009). *Policy and Procedures Manual*. Australian Capital Territory, Canberra. Retrieved from http://www.bsss.act.edu.au/publications/policies_and_prodedures

Ainley, J., Fleming, M., & McGregor, M. (2002). *Three years on- literacy advance in the early and middle primary years*. Melbourne, Australia: Catholic Education Commission of Victoria.

Alagumalai, S., Keeves, J.P., & Hungi, N. (1996). *Disattenuated Correlation and Unidimensionality*. Paper presented to Australian Association for Research in Education. Retrieved from http://www.aare.edu.au/96pap/alags96453.txt

Alexander, K., Entwisle, D., & Olson, L. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis, 23*(2), 171-191.

Allen, M.J.& Yen, W.M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Altman, D.G. (1991). *Practical statistics for medical research*. London, England: Chapman and Hall.

Andersen, E.B., & Olsen, L.W. (2001). The life of Georg Rasch. In A. Boomsma, M. van Duijn, T. Snijders (Eds.). *Essays on Item Response Theory*. New York, NY: Springer.

Arnold, M. (1889). *Reports on Elementary Schools 1852-1882*. London, England: MacMillan.

Assessment and Evaluation Unit, University of Leeds. (2004). *Evaluation of the trial assessment arrangements for key stage 1:Report to QCA*. The University of Leeds Assessment and Evaluation Unit. Retrieved from http://www.qca.org.uk/qca_10082.aspx.

Assessment Reform Group-ARG. (2006). *The role of teachers in the assessment of learning*. Retrieved from http://k1.ioe.ac.uk/tlrp/arg/index.html

Aunola, K., Leskinen, E., Lerkkanen, M., & Nurmi, J. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology, 96*, 699–713.

Ayres, L.P. (1912). A Scale for Measuring the Quality of Handwriting of School Children. New York, NY: Russell Sage Foundation.

Ayres, L.P. (1918). History and present status of educational measurements. *Seventeenth Yearbook of the National Society for the Study of Education*, *17*(2) 11-12. Cited in Cadenhead, K, & Robinson, R (1987). Fisher's "Scale-Book": An early attempt at educational measurement. Educational Measurement: Issues and Practice, 6(4), 15-17.

Bangert-Drowns, R.L., Kulik, C.C., Kulik, J.A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*, 213-238.

Barnhart, H.X., Haber, M.J.& Lin, L.I. (2007). An overview on assessing agreement with continuous measurements, *Journal of Biopharmaceutical Statistics*, *17*, 529-569.

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology, 21*(2), 177-187.

Bedard,K., & Dhuey, E. (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *The Quarterly Journal Of Economics, 121*, 1437-1472.

Bennett, R.E., Gottesman, R.L., Rock, D.A., & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgements of students' academic skill. *Journal of Educational Psychology, 85*, 347–356.

Biggs, J., & Collis, K. (1982). *Evaluating the Quality of Learning: the SOLO Taxonomy*. New York, NY: Academic Press.

Black, P. (2003). (with the King's College London Assessment for Learning Group Harrison, C., Lee, C., Marshall, B., Williams, D.). *Formative and summative Assessment: Can They Serve Learning Together?* Paper presented at American Educational Research Association Classroom Assessment SIG. Retrieved from http://www.kcl.ac.uk//depsta/education/hpages/pblackpubs.html

Black, P.J.; Harrison, C.; Lee, C.; Marshall, B., & Wiliam, D (2002). *Working inside the black box: Assessment for learning in the classroom.* London, England: King's College London School of Education.

Black, P., & Wiliam, D. (1998a). Assessment and Classroom Learning. *Assessment in Education, 5*, 7-74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139-148.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: putting it into practice*. Buckingham, England: Open University Press.

Block, J.H. (Ed.). (1971). *Mastery learning: Theory and practice*. New York, NY: Holt, Rinehart & Winston.

Bloom, B.S. (1968). Learning for mastery. *Evaluation Comment*, *1*, 2, 1–12. Retrieved from ERIC database. (ED053419)

Bloom, B.S. (1971). Mastery learning. In J.H. Block (Ed.), *Mastery learning: Theory and practice* (p. 47–63). New York, NY: Holt, Rinehart & Winston.

Bond, T.G., & Bond, M.L. (2003, November). *Measure for measure: Curriculum requirements and children's achievement in music education.* Paper presented at the Annual Conference of the New Zealand and Australian Associations for Research in Education, Auckland, New Zealand.

Bond, T.G., & Fox, C.M. (2001). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: Erlbaum.

Bond, T.G., & Fox, C.M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences. (2nd ed.) Mahwah, NJ: Erlbaum.

Boomer, G. (1990). *Attainment levels: The starting orders*, unpublished internal directive to members of the South Australian Education Department's Directorate of Curriculum cited by Jenkin (1996).

Boston, K. (1992). Working for the possible. *Curriculum Perspectives*, *12*(4), p. 15-16.

Boston, K. (1993). Introducing profiles. *Curriculum Perspectives*, *13*(2), p. 12-13.

Boston, K. (1994). A perspective on the so-called 'National Curriculum'. *Curriculum Perspectives*, *14*(1), p. 43-45.

Brookhart, S.M. (2004). Classroom assessment: tensions and intersections in theory and practice. *Teachers College Record*, *106*, 429-458.

Brown, C.R., Moor, J.L., Silkstone, B.E., & Botton, C. (1996). The construct validity and context dependency of teacher assessment of practical skills in some pre-university level science examinations. *Assessment in Education, 3*, 377–391.

Brown, C.R., Moor, J.L., Silkstone, B.E., & Botton, C. (1998). An evaluation of two different methods of assessing independent investigations in an operational pre-university level examination in biology in England. *Studies in Educational Evaluation, 24*, 87–98.

Burgess, M.A. (1937). The construction of two height charts. *Journal of the American Statistical Association, 32*, 198, 290-310.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology*, *58*, 1-14.

Butts, R.F.(1978). Public Education in the United States: From Revolution to Reform. New York, NY: Holt, Rinehart and Winston.

Cadenhead, K, & Robinson, R (1987). Fisher's "Scale-Book": An early attempt at educational measurement. *Educational Measurement: Issues and Practice, 6*(4), 15-17.

Cahan, S., & Davis, D. (1987). A between-grade-levels approach to the investigation of the absolute effects of schooling on achievement. *American Educational Research Journal, 24*, 1-12.

Caldwell, O.W., & Courtis, S.A. (1925). Then & now in education 1845:1923: A message of encouragement from the past to the present. Yonkers-On-Hudson, NY: World Book Company.

Callingham, R. (2003, November). *Establishing the validity of a performance assessment in numeracy.* Paper presented to the Australian Association for Research in Education Conference, Auckland. Retrieved from http://www.aare.edu.au/03pap/cal03240.pdf

Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. *Journal of Educational Measurement*, *36*, 73–78.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*(4), 379-388.

Canto, A.I. (2006). Predicting third grade students' FCAT reading achievement and oral reading fluency using student demographic, academic history, and performance indicators. Ph D dissertation, Florida State University. Retrieved from Etd.Lib.Fsu.Edu/Theses/Available/Etd-04102006-114229/Unrestricted/ Dissertation_Angel_Canto.pdf

Care, E., Griffin, P., Thomas, A., Pavlovic, M. (2007). *Early years assessment of English*. Melbourne, Australia: Assessment Research Centre, University of Melbourne.

Caroline Springs College (2008). *Annual Report to the School Community*. Retrieved from www.carolinesprings.vic.edu.au

Center for Disease Control (2000). *2 to 20 years: Girls/boys stature weight-for-age percentiles*. Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion. Retrieved from http://www.cdc.gov/growthchart

Chadwick, E.B.(1864). The Museum, a Quarterly Magazine of Education, Literature and Science. Vol. II. Reprinted in the *Journal of the Statistical Society of London*, *27*(2) (Jun., 1864), 261-266.

Clay, M.M. (1972). *Reading: the patterning of complex behaviour*. Auckland, New Zealand: Heinemann Educational Books.

Code of Conduct for Using Student Achievement Information (1995). *An Appendix to the assessment and reporting for schools policy statement*. Adelaide, Australia: Department of Education and Children's Services.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78,* 141–146.

Collins, C. (1994a). Curriculum and pseudo-science: Is the Australian national curriculum project built on credible foundations? *Occasional Paper No. 2*. Canberra: Australian Curriculum Studies Association.

Collins, C. (1994b). Is the National Curriculum Profiles brief valid? *Curriculum Perspectives*, *14(*1), p. 45-48.

Collins, L.M.(2006). Analysis of longitudinal data: The integration of theoretical models, temporal design, and statistical models. *Annual Review of Psychology*, *57*, 505–528.

Condry, K.F., & Spelke, E.S. (2008). The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology, 137*, 22–38.

Cooper, H., & Tom, D. (1984). Teacher expectation research: A review with implications for classroom instruction. *The Elementary School Journal, 85*, 76-89.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*, 227-268.

Courtis, S.A. (1909a). Measurement of growth and efficiency in arithmetic. *The Elementary School Teacher*, *10*, 58-74.

Courtis, S.A. (1909b). Measurement of growth and efficiency in arithmetic (continued). *The Elementary School Teacher*, *10*, 177-199.

Courtis, S.A. (1910). Measurement of growth and efficiency in arithmetic (continued). *The Elementary School Teacher*, *11*, 171-185.

Courtis, S.A. (1911a). Measurement of growth and efficiency in arithmetic (continued). *The Elementary School Teacher*, *11*, 360-370.

Courtis, S.A. (1911b). Measurement of growth and efficiency in arithmetic (continued). *The Elementary School Teacher*, *11*, 528-539.

Courtis, S.A. (1911c). Standard scores in arithmetic. *The Elementary School Teacher*, *12*, 127-137.

Courtis, S.A. (1913). The reliability of single measurements with standard tests. *The Elementary School Teacher*, *13*, 326-345.

Courtis, S.A. (1914). Standard tests in English. *The Elementary School Teacher*, *14*, 374-392.

Courtis, S.A. (1916). Courtis tests in arithmetic: Value to superintendents and teachers. *The Fifteenth Yearbook of the National Society For The Study of Education. (*pp. 91-106). Chicago, IL: University of Chicago Press, .

Courtis, S.A. (1917). *Courtis standard practice tests in arithmetic: Teacher's manual*. Yonkers-On-Hudson, NY: World Book Company.

Courtis, S.A. (1919). The uses of the Hillegas scale. *The English Journal, 8*(4), 203-217.

Courtis, S.A. (1929). Maturation units for the measurement of growth. *School and Society*, *30*, 683-690.

Crawford, C., Dearden, L., & Meghir, C. (2007). When you are born matters: The impact of date of birth on child cognitive outcomes in England. The Institute for Fiscal Studies. Retrieved from http://www.ifs.org.uk/publications.php?publication_id=4073

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*, 438-481.

Cross, A. (1917). Weighing the scales. *The English Journal, 6*(3), 183-191.

Cudeck, R., & Harring, J.R. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology, 58*, 616–637.

Cumming, J.J., Wyatt-Smith, C., Elkins, J., & Neville, M. (2006). *Teacher Judgment: Building An Evidentiary Base For Quality Literacy And Numeracy.* Griffith University. Retrieved from www.qsa.qld.edu.au/downloads/publications/research_qsa_teacher_judgment.pdf

Curriculum Corporation. (1994a). *The statements and profiles for Australian schools*. (16 vols.) Melbourne, Australia: Author.

Curriculum Corporation. (1994b). *CURASS guidelines papers,* Melbourne, Australia: Author.

Curriculum Corporation. (1994c). *English –a curriculum profile for Australian schools*. Melbourne, Australia: Author.

Darwin, C. (1860). A naturalist's voyage round the world: Journal of researches into the natural history and geology of the countries visited during the voyage round the world of H.M.S. Beagle under the command of Captain Fitz Roy, R.N. (1913 edition). London, England: John Murray.

Daugherty, R. (1997). National curriculum assessment: The experience of England and Wales. *Educational Administration Quarterly*, *33*, 198-218.

De Ayala, R.J. (2008). A commentary on historical perspectives on invariant measurement: Guttman, Rasch and Mokken. *Measurement: Interdisciplinary Research and Perspectives*, *6*(3), 209-212.

Demaray, M.K., & Elliot, S.N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13,* 8–24.

Department for Children, Schools and Families. (2009). *The national strategies: AfL with APP: Developing collaborative school-based approaches: Guidance for senior leaders*. Retrieved from http://nationalstrategies.standards.dcsf.gov.uk/.

Department of Education and Early Childhood Development, Victoria. (2003). *Years Prep - 10 Curriculum and Standards Framework II - Benchmarks, 2002*. Spreadsheet CSF_BENCHMARKS_02_(v1.1)-1.xls. Retrieved from http://www.education.vic.gov.au/management/schoolimprovement/performancedata/publications.htm

Department of Education and Early Childhood Development, Victoria. (2006). *Years Prep - 10 Curriculum and Standards Framework II - Benchmarks, 2005*. Spreadsheet CSF_BENCHMARKS_05_(v1.0).xls. Retrieved from http://www.education.vic.gov.au/management/schoolimprovement/performancedata/publications.htm

Department of Education and Early Childhood Development, Victoria. (2009a). *English Online Interview School User Guide.* Retrieved from http://www.education.vic.gov.au/studentlearning/teachingresources/english/englishonline.htm#2

Department of Education and Early Childhood Development, Victoria. (2009b). *Mathematics Online Interview School User Guide*. Retrieved from https://www.eduweb.vic.gov.au/MathematicsOnline/

Department of Education, Employment and Training, Victoria. (1996). *Assessment and Reporting Support Materials*. Retrieved August 2004 from http://www.sofweb.vic.edu.au

Department of Education and Training, Victoria. (2002). Teacher Judgement Benchmarks 2001: *Years Prep-10 CSF Benchmarks*. Standards and Accountability Division, Office of School Education, Department of Education and Training. Victoria. Retrieved from http://www.education.vic.gov.au/management/schoolimprovement/performancedata/publications.htm

Department of Education, Victoria. (1997). *Benchmarks 96*: *Years Prep–10: curriculum and standards framework.* Retrieved from http://www.education.vic.gov.au/management/schoolimprovement/performancedata/publications.htm

Department of Education, Victoria. (1998). *Benchmarks 97*: *Years Prep–10: curriculum and standards framework.* Retrieved from http://www.education.vic.gov.au/management/schoolimprovement/performancedata/publications.htm

Department of Education, Victoria. (1999). *Years Prep –10 curriculum and standards framework benchmarks 1998*. Office of Review. Retrieved from www.education.vic.gov.au/management/schoolimprovement/performancedata/publications.htm

Donnelly, K. (2007). Australia's adoption of outcomes based education. *Issues In Educational Research*, *17*. Retrieved from http://www.iier.org.au/iier17/donnelly.html

Du Bois, P.H. (1970). *A history of psychological testing*. Boston, MA: Allyn & Bacon.

Duncan, A (2009). *Teacher preparation: Reforming the uncertain profession*—Remarks of Secretary Arne Duncan at Teachers College, Columbia University. October 22. Retrieved from http://www.ed.gov/print/news/speeches/2009/10/10222009.html

Dunn, G. (2007). Regression models for method comparison data, *Journal of Biopharmaceutical Statistics*, *17*, 739-756.

Dunn, L., Morgan, C., O'Reilly, M., & Parry, S. (2004). *The student assessment handbook.* London, England: Routledge & Falmer.

Durant, D. (2003, September). *A comparative analysis of Key Stage tests and teacher assessments*. Paper presented to the British Education Research Association Annual Conference. Retrieved from www.leeds.ac.uk/educol/documents/00003153.htm

Early Childhood Longitudinal Study. (2004). *Data files and electronic code book*. US Dept of Education, Institute of Education Sciences. NCES 2004-089.

Education Department of South Australia (1992). *Attainment levels.* Adelaide, Australia: Author.

Education Department of South Australia (1993). Monitoring Student Achievement Resource Paper 2: Attainment Levels and National Profiles. Adelaide, Australia: Author.

Egan, O., & Archer, P. (1985). The accuracy of teachers' ratings of ability: A regression model. *American Educational Research Journal*, *22*, 25-34.

Endler, L.C., & Bond, T.G. (2008).Changing science outcomes: Cognitive acceleration in a US setting. *Research in Science Education, 38*(2), 149-166.

Engelhard, G. (1984). Thorndike, Thurstone, and Rasch: A comparison of their methods of scaling psychological tests. *Applied Psychological Measurement*, *8*, 21-38.

Engelhard, G. (1991a). Thorndike, Thurstone, and Rasch: A comparison of their approaches to item-invariant measurement. *Journal of Research and Development in Education*, *24*(2), 45-60.

Engelhard, G. (1991b). Thorndike and Wood. *Rasch Measurement Transactions*, *5*(2), 146.

Engelhard, G. (1992a). Historical views of invariance: Evidence from the measurement theories of Thorndike, Thurstone, and Rasch. *Educational and Psychological Measurement*, *52*, 275-291.

Engelhard, G. (1992b). Thorndike's scaling vs. Wood's scoring. *Rasch Measurement Transactions*, *5*(4), 182.

Farley, T. (2009). Making the grades: My misadventures in the standardized testing industry. Sausalito, CA: PoliPointPress.

Feinberg, A.B.& Shapiro, E.S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly*, *18*, 52–65.

Fisher, G. (1862). *On the numerical mode of estimating educational qualifications, as pursued at the Greenwich Hospital School*. Paper presented to 32nd Meeting of the British Association for the Advancement of Science, Section F –Economic Science and Statistics, October 1862, listed in the Journal of the Statistical Society of London, *25*(4), Republished by Cadenhead & Robinson, 1987.

Fitz-Gibbon, C.T. (1992). Peer and cross-age tutoring. In M.C. Alkin (Ed.) *Encylopedia of Educational Research*. (pp. 980-984). New York, NY: Macmillan.

Forster, M. (2009, August). *Informative assessment: Understanding and guiding learning*. Paper presented to the Australian Council for Educational Research Assessment and Student Learning Conference, Perth.

Forster, M., & Masters, G. (1996). *Portfolios assessment resource kit (ARK Portfolios)*. Melbourne, Australia: Australian Council for Educational Research.

Forster, M., & Masters, G. (2004). Bridging the conceptual gap between classroom assessment and system accountability. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability.* 103rd Yearbook of the National Society for the Study of Education. (pp. 51-73). Chicago, IL: University of Chicago Press.

Frederiksen, J.R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Freund, P.A., Holling, H., & Preckel, F. (2007). A multivariate, multilevel analysis of the relationship between cognitive abilities and scholastic achievement. *Journal of Individual Differences, 28*(4), 188-197.

Frigo, T. (1997, November). *Using curriculum statements and profiles in South Australian schools*. Paper presented at the AARE Annual Conference, Brisbane. Retrieved from www.aare.edu.au/97pap/frigt213.htm

Frigo, T. (1998). Refining the curriculum statements and profiles: report of focus group sessions with teachers. Melbourne, Australia: Australian Council for Educational Research.

Fuchs, L.S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*, 199-208.

Fullan, M., Hill, P., & Crevola, C. (2006). *Breakthrough*. Thousand Oaks, CA: Corwin.

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 American Educational Research Association Annual Meeting, April 2000, Retrieved from ERIC database. (ED 441015).

Garner, R. (2009, April 12). Chanting teachers welcome vote to boycott primary tests. *The Independent.* Retrieved from

http://www.independent.co.uk/news/education/education-news/chanting-teachers-welcome-vote-to-boycott-primary-tests-1667517.html

Garrett, R., & Plitz, W. (1999). A case study of curriculum control: curriculum reform in health and physical education. In B. Johnson, & A. Reid (Eds.). *Contesting the curriculum*. Katoomba, Australia: Social Science Press.

Gilmore, J., & Vance, M. (2007). Teacher ratings of children's listening difficulties. *Child Language Teaching and Therapy, 23*, 133-156.

Gipps, C., Brown, M, McCullum, B., & McAlister, S. (1995). *Intuition or evidence? Teachers and national assessments of seven-year-olds.* Buckingham, England: Open University Press.

Gipps, C.V., Murphy, P. (1994). *A fair test?: Assessment, achievement and equity*. Buckingham, England: Open University Press.

Gladwell, M. (2005). *Blink: The power of thinking without thinking*. New York, NY: Little, Brown and Co.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality. *Philosophical Transactions of the Royal Society*, *115*, 513-583. Retrieved from http://rstl.royalsocietypublishing.org.elibrary.jcu.edu.au/content/115.toc

Good, H.G. (1926). An early school surveyor, *Educational Research Bulletin*, *5*, 351-353.

Gough, J. (2006). VELS versus CSF; replacement, conflict or harmony? *Vinculum*, *43*(2), 3-6.

Green, S.K., & Mantz, M. (2002). *Classroom assessment practices: Examining impact on student learning.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. ED 464 920.

Griffin, P.E. (1990). Profiling literacy development: Monitoring the accumulation of reading skills. *Australian Journal of Education*, *34*, 290-311.

Griffin, P.E. (1998). Outcomes and profiles: Changes in teachers' assessment practices. *Curriculum Perspectives, 18*(1), 9-19.

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

Griffin, P.E. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation*, *33*, 87-99.

Griffin, P., & Callingham, R. (2006). A twenty-year study of mathematics achievement. *Journal for Research in Mathematics Education, 37*(3), 167-186.

Grissom, J.B. (2004). Age and achievement. *Education Policy Analysis Archives*, *12*(49), 1-40. Retrieved from http://epaa.asu.edu/ojs/article/view/204

Gunther, C. (1919). My experience with the Hillegas scale. *The English Journal*, *8*, 535-542.

Guskey, T.R. (1981). Comparison of a Rasch model scale and the grade-equivalent scale for vertical equating of test scores. *Applied Psychological Measurement*, *5*, 187-201.

Gustafsson, J. (1979). The Rasch Model in Vertical Equating of Tests: A Critique of Slinde and Linn. *Journal of Educational Measurement*, 16, 3, 153-158.

Haertel, E. (1991). Report on TRP analyses of issues concerning within-age versus cross-age scales for the National Assessment of Educational Progress. Washington, DC: National Center for Education Statistics. Retrieved from ERIC database. (ED404367)

Haertel, E., & Herman, J. (2005). *A Historical Perspective on Validity Arguments for Accountability Testing.* CSE Report No. 654.. Technical Report for the National Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Hampton Primary School (2006). *Sample whole year reporting schedule and extracts from whole year assessment calendars*. Retrieved from www.education.vic.gov.au/studentlearning/studentreports/casestudies/default.htm

Harlen, W. (1994). Towards quality in assessment, in W. Harlen (Ed.), *Enhancing quality in assessment*. London, England: Paul Chapman.

Harlen, W. (2004a). A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes. London: EPPI-Centre, Social Science Research Unit, Institute of Education. Retrieved from http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=149

Harlen, W. (2004b). *A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes.* London: EPPI-Centre, Social Science Research Unit, Institute of Education. Retrieved from http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=149

Harlen, W. (2005a). Teachers' summative practices and assessment for learning-tensions and synergies. *The Curriculum Journal, 16*, 207-223.

Harlen, W. (2005b). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*, 245–270.

Harlen, W. (2007a). Criteria for evaluating systems for student assessment. *Studies In Educational Evaluation, 33*, 15–28.

Harlen, W. (2007b). The quality of learning: assessment alternatives for primary education. *Primary Review Research Survey 3/4*, Cambridge, England: University of Cambridge Faculty of Education.

Harlen, W. (2007c). *Assessment of learning*. London, England: Sage.

Hattie, J. (1999). *Influences on student learning:. inaugural professorial address*. August 2, 1999. Auckland, New Zealand: University of Auckland. Retrieved from www.arts.auckland.ac.nz/edu/staff/jhattie

Hattie, J. (2003, October). *Teachers make a difference: What is the research evidence?* Paper presented at the Australian Council for Educational Research Annual Conference on Building Teacher Quality, Melbourne.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Hattie, J.A., & Brown, G.T. (2003). Standard setting for asTTle reading: A comparison of methods. *asTTle Technical Report 21*, University of Auckland/Ministry of Education. Retrieved from http://www.tki.org.nz/r/asttle/tech-reports_e.php

Hauser, C. (2003, April). *So, what d'ya expect? Pursuing reasonable individual student growth targets to improve accountability systems*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago. Retrieved from http://www.kingsburycenter.org/

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Paper prepared for the formative assessment for teachers and students. State Collaborative on Assessment and Student Standards of The Council Of Chief State School Officers. Washington DC.

Herman, J., & Choi, K. (2008). Formative assessment and the improvement of middle school science learning: The role of teacher accuracy. *CRESST Report 740*. Retrieved from www.cse.ucla.edu/products/reports/R740.pdf

Hieronymus, A.N. & Hoover, H.D. (1986). *Iowa Tests of Basic Skills manual for school administrators*. Chicago: Riverside.

Hill, C.J., Bloom, H.S., Rebeck Black, A., & Lipsey, M.W. (2007). Empirical benchmarks for interpreting effect sizes in research. *MDRC Working Papers on Research Methodology*. Retrieved from www.mdrc.org

Hill, P.W. (1994). Putting the national profiles to use. *Unicorn*, *20*(2), 36-42.

Hillegas, M. (1912). Scale for the measurement of quality in English composition by young people. *Teachers College Record*, *13*(4), 1-55.

Hilton, T.L. & Patrick, C. (1970). Cross-sectional versus longitudinal data: an empirical comparison of mean differences in academic growth. *Journal of Educational Measurement*, *7*, 1,15-24.

Hinnant, J.B., O'Brien, M., & Ghazarian, S.R. (2009). The longitudinal relations of teacher expectations to achievement in the early years. *Journal of Educational Psychology*, *101*, 662-670.

Hinton, E.M. (1940). An analytical study of the qualities of style and rhetoric found in English compositions. *Teachers College Record*, *42*(2), 157-159. Retrieved from http://www.tcrecord.org (ID Number: 8975)

Hoeksma, J.B., & Kelderman, H. (2006). Commentary: On growth curves and mixture models. *Infant and Child Development*, *15*, 627–634.

Hoff, D.J. (2006). Delving into data. *Education Week Technology Counts 2006*. *25*(5), 12-14 and 20-22. Retrieved from http://www.edweek.org/

Hoge, R.D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research, 59,* 297–781.

Holmes. (1913). From Professor Holmes. *The English Leaflet, 104*, 4-5. Retrieved from http://www.archive.org/details/hillegasscale00newerich

Holmes, S.E. (1982). Unidimensionality and vertical equating with the Rasch model, *Journal of Educational Measurement, 19*, 2, 139-147.

Hornibrook, M., & Wallace, M. (2001). Report on the Independent Evaluation of the Development of the South Australian Curriculum Standards and Accountability Framework. Melbourne, Australia: Curriculum Corporation. Retrieved from http://www.sacsa.sa.edu.au

Hubbard, D.W. (2007). How to measure anything: Finding the value of intangibles in business. New York, NY: John Wiley and Sons.

Hudelson, E. (1916). Some achievements in the establishment of a standard for the measurement of English composition in the Bloomington, Indiana, schools. *The English Journal*, *5*, 590-597.

Hudelson, E. (1923). The development and comparative values of composition scales. *The English Journal*, *12*(3), 163-168.

Humboldt, A. von, (1811). Atlas geographique et physique du royaume de la nouvelle-espagne, fonde sur des observations astronomiques, des mesures trigonometriques et des nivellemens barometriques. No. 29. Paris: F. Schoell. Cited in H.G Funkhouser (1937). Historical development of the graphical representation of statistical data. *Osiris, 3*, 269-404. Saint Catherines Press. Retrieved from http://www.jstor.org/stable/301591

Hungi, N. (2003). *Measuring school effects across grades*. Institute of International Education Research Collection, Number 6. Adelaide, Australia: Flinders University.

Huxley, J. (Ed.). (1936). *T.H. Huxley's diary of the voyage of H.M.S. Rattlesnake*. Garden City, NY: Doubleday, Doran and Company.

Hyams, D. (2001). *CurveExpert Version 1.38 A curve fitting system for Windows*. Copyright 1995-2001. Retrieved from http://curveexpert.webhop.net/

Jenkin, R. (1996). *Australian national collaborative curriculum development- a tangled web*. Contribution to M. Dilena and C.E. van Kraayenoord website: Whole school approaches to assessing and reporting literacy. Retrieved 27 September 2004 (link now broken) from http://www.gu.edu.au/school/cls/clearinghouse/

Johanningmeier, E.V. (2004). The transformation of Stuart Appleton Courtis: Test maker and progressive, *American Educational History Journal*, *31*(2), 202-210.

Johanningmeier, E.V. & Richardson, T.R. (2008). *Educational Research, The National Agenda, and Educational Reform: A History.* Charlotte, NC: Information Age Publishing.

Johnson, B., & Reid, A. (Eds.). (1999). *Contesting the curriculum*. Katoomba, Australia: Social Science Press.

Johnson, F.W. (1913). The Hillegas-Thorndike scale for measurement of quality in English composition by young people. *The School Review*, *21*(1), 39-49.

Johnson, J.A., Dupuis, V.L., Musial, D., Hall, G.E., & Gollnick, D. (2002). *Introduction to the foundations of American Education* (12th ed.). Boston: Allyn, & Bacon.

Jorgensen, M.A. (2004). And there is much left to do. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability.* 103rd Yearbook of the National Society for the Study of Education. (pp. 203-208). Chicago, IL: University of Chicago Press.

Jussim, L., & Eccles, J. (1995). Naturalistic studies of interpersonal expectancies. *Review of Personality and Social Psychology*, *15*, 74–108.

Justice, L.M., Pence, K., Bowles, R.B., & Wiggins, A. (2006). An investigation of four hypotheses concerning the order by which 4-year-old children learn the alphabet letters. *Early Childhood Research Quarterl, 21*, 374–389.

Keats, J.A. (1983). Ability measures and theories of cognitive development. In H. Wainer & S. Messick (Eds.) *Principles of modern psychological measurement: A festschrift for Frederick M. Lord*. Hillsdale, NJ: Erlbaum.

Keeves, J.P. (Chair). (1982). Education and change in South Australia: Final Report of the Committee of Enquiry into Education in South Australia. Adelaide, Australia: Government Printer.

Keeves, J.P, & Marjoribanks, K. (Eds.). (1999). *Australian education: Review of research 1965-1998*. Melbourne, Australia: Australian Council for Educational Research.

Kerbow, D., & Bryk, A. (2005). *STEP literacy assessment technical report of validity and reliability*. Center for Urban School Improvement (USI) at the University of Chicago. Retrieved from http://www.iisrd.org/software_step/index.shtml

Kingsbury, G.G.(2003). *A Long-Term Study Of Item Parameter Estimates*. Paper presented to the Annual Meeting of the American Educational Research Association, April, Chicago.

Kissane, B.V. (1982). The Measurement of Change as the Study of the Rate of Change. *Education Research and Perspectives*, *9*(1), 55-72.

Klein, G. (1999). *Sources of power. How people make decisions*. Cambridge, MA: MIT Press.

Klein, G. (2009). Streetlights and shadows. Searching for the keys to adaptive decision making. Cambridge, MA: MIT Press.

Kolen, M.J., & Brennan, R.L.(2004). *Test equating, scaling, and linking-Methods and practices*. (2nd ed.). New York, NY: Springer-Verlag.

Laidra, K., Allik, J., Harro, M., Merenäkk, L., & Harro, J. (2006). Agreement among adolescents, parents, and teachers on adolescent personality. *Assessment, 13*(2), 187-196.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Leahy, S., & Wiliam, D. (2009). *From teachers to schools: scaling up professional development for formative assessment*. Paper presented at American Educational Research Association conference, Retrieved from http://www.dylanwiliam.net/

Learned, W.S. (1913). From Dr. Learned. *The English Leaflet, 104*, 7-8. Retrieved from http://www.archive.org/details/hillegasscale00newerich

Lee, O.K. (1993). Absolute zeroes of reading and mathematics. *Rasch Measurement Transactions*, *6*(4), 245-246.

Lee, O.K. (2003*),* Rasch simultaneous vertical equating for measuring reading growth. *Journal of Applied Measurement*, *4*(1), 10-23.

Ligon, G.D. (2009). Growth models – finding real gains. Extraordinary insight into today's education information topics. Retrieved from http://www.espsolutionsgroup.com

Linacre J.M., & Wright B.D (1989). The "length" of a logit. *Rasch Measurement Transactions, 3*(2), 54-55.

Linacre, J.M. (1998). Wright: The measure of the man. *Popular Measurement*, *Spring*, 23-25.

Linacre, J.M. (1999). Explorations into local independence with T-Rasch. *Rasch Measurement Transactions*, *13*(3), 710.

Linacre, J.M. (2006). A User's Guide to WINSTEPS MINISTEP Rasch-model computer programs. Retrieved from www.winsteps.com

Lokan, J. (Ed) (1997). *Describing Learning; Implementation of Curriculum Profiles in Australian Schools 1986-1996*. Melbourne, Australia: Australian Council for Educational Research.

Lokan, J., & Wu, M. (1997). Summary of the 1992-93 ACER calibration studies. In J. Lokan (Ed.). (1997). *Describing Learning; Implementation of Curriculum Profiles in Australian Schools 1986-1996. (*pp. 8-23). Melbourne, Australia: Australian Council for Educational Research.

Macken, E., Suppes, P., & Zanotti, M. (1980). Considerations in evaluating individualised instruction. *Journal of Research and Development in Education, 14*(1), 79-83.

Malone, T.W., Suppes, P., Macken, E., Zanotti, M., & Kanerva, L. (1979). Projecting student trajectories in a computer-assisted instruction curriculum. *Journal of Educational Psychology*, *71*, 74-84.

Mann, H. (1845). Boston Grammar and Writing Schools. *The Common School Journal*. Reprinted in O.W. Caldwell & S.A. Courtis. (1925). *Then & now in education 1845:1923: A message of encouragement from the past to the present. (*pp. 237-272). Yonkers-On-Hudson, NY: World Book Company.

Mantel, H. (2005), Is the particle there? Review of *A Game with Sharpened Knives* by Neil Belton. *London Review of Books, 27*(13), 13-14.

Marist-Sion College. (2008). *School Annual Report*. Retrieved from http://www.vrqa.vic.gov.au/.

Marsh, C.J. (1994). *Producing a national curriculum: Plans and paranoia*. Sydney Australia: Allen and Unwin.

Marsh, J.A., Pane, J.F., & Hamilton, L.S. (2006). *Making sense of data-driven decision making in education: evidence from recent RAND research*. Retrieved from http://www.rand.org/pubs/occasional_papers/OP170/

Marzano, R. (2000a). *Transforming classroom grading*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. (2000b). *Designing a new taxonomy of educational objectives*. Experts In Assessment Series. Thousand Oaks, CA: Corwin Press.

Marzano, R.J. (2007). *The art and science of teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Masters, G.N. (1990). *Subject profiles as frameworks for assessing and reporting student achievement*. Unpublished discussion paper, commissioned by the Australasian Cooperative Assessment Program, Australian Council for Educational Research, Melbourne.

Masters, G.N. (1992). Towards a national framework for assessment and reporting. *Unicorn, 18*(1), 66-77.

Masters, G.N. (1999). Research in educational measurement. In J. Keeves & K. Majoribanks (Eds.). (1999). *Australian education: Review of research 1965-1998*. Melbourne, Australia: Australian Council for Educational Research.

Masters, G.N. (2005). Against the grade: In search of continuity in schooling and learning. *Professional Educator*, *4*(1), 12-22.

Masters, G.N., & Forster, M. (1996). *Developmental assessment: Assessment resource kit. (*ARK developmental assessment). Melbourne, Australia: Australian Council for Educational Research.

Masters, G.N., & Forster, M. (1996). *Progress maps: Assessment resource kit*. Melbourne, Australia: Australian Council for Educational Research.

Masters, G.N., & Forster, M. (1997). *Mapping literacy achievement, results of the 1996 national school English literacy survey.* Canberra, Australia: Commonwealth of Australia. Retrieved from http://www.dest.gov.au/mla/mla.pdf

Masters, G.N., & Forster, M.(undated). *The assessments we need.* Australian Council for Educational Research website. Retrieved from www.acer.edu.au/research/ documents/Theassessmentsweneed.pdf

Masters, G.N., & Keeves, J.P. (Eds.). (1999). Advances in measurement in educational research and assessment. New York, NY: Pergamon (Elsevier Science).

Masters, G. N., Lokan, J., Doig, B., Khoo, S.T., Lindsey, J., Robinson, L. & Zammit, S. (1990). *Profiles of Learning: The Basic Skills Testing Program in New South Wales 1989.* Melbourne, Australia: Australian Council for Educational Research. Retrieved from ERIC database. (ED327276)

Masters, G.N., & Mossenson, L. T. (1983). *Using a Latent Trait Model to Vertically Equate Educational Tests*. Paper presented at Australian Association for Research in Education Conference, Canberra.

Masters, G.N., Rowley, G., Ainley, & J. Khoo, S. T. (2008). *Reporting and Comparing School Performances. Paper prepared for the MCEETYA Expert Working*. Commissioned by the Reporting and Accountability Branch, National Education

Systems Group, Commonwealth Department of Education, Employment and Workplace Relations (DEEWR).

Maxwell, G.S. (2004). *Progressive assessment for learning and certification: some lessons from school-based assessment in Queensland*. Paper presented at the third conference of the Association of Commonwealth Examination and Assessment Boards, March, Nadi, Fiji. Retrieved from http://www.spbea.org.fj/aceab_conference.html

Mazyck, M. (2002). Integrated learning systems and students of color: Two decades of use in K-12 education. *TechTrends, 46*(2), 33.

McCall, M., Hauser, C., Cronin, J., Kingsbury, G., & Houser, R. (2006). Achievement gaps: An examination of differences in student achievement and growth. Technical report from the NWEA Growth Research Database. Retrieved from http://www.kingsburycenter.org/our-research/research-reports-publications

McCall, M.S. (2006). *Item Response Theory And Longitudinal Modeling: The Real World Is Less Complicated Than We Fear*. Presented to the MSDE/MARCES Conference Assessing & Modeling Cognitive Development In School: Intellectual Growth And Standard Setting October. [PowerPoint slides] Retrieved from http://www.education.umd.edu/EDMS/MARCES/conference/cognitive/MARCESPresMcCalltiny.ppt

McCall, W.A. (1922). *How to Measure in Education.* New York: Macmillan.

McGaw, B. (1994). Standards from a curriculum and assessment perspective. *Queensland Researcher, 10*(2), 1-18. Retrieved from http://education.curtin.edu.au/iier/qjer/qr10/mcgaw.html

McKinsey & Company. (2007). *How the World's Best-performing School Systems Come Out on Top*. Retrieved from http://www.mckinsey.com/clientservice/socialsector/resources/pdf/Worlds_School_Systems_Final.pdf

Meisels, S., Bickel, D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment from Kindergarten to Grade 3. *American Educational Research Journal; 38*, 73-95.

Meisels, S., Dorfman, A., & Steele, D. (1994). *Equity and excellence in group-administered and performance-based assessments*. In M. Nettles & A. Nettles (Eds.), *Equity in educational assessment and testing* (pp. 195—211). Boston, MA: Kluwer Academic Publishers.

Meisels, S.J., Liaw, F., Dorfman, A., & Nelson, R. (1995). The Work Sampling System: reliability and validity of a performance assessment for young children. *Early Childhood Research Quarterly, 10*, 277-296.

Merrell, C., & Tymms, P. (2007). What children know and can do when they start school and how this varies between countries. *Journal of Early Childhood Research, 5*, 115-134.

Merton, R. (1948). The self-fulfilling prophecy. *Antioch Review*, *8*, 193-210.

Messick, S. (1993). *Foundations of validity: Meaning and consequences in psychological assessment.* Princeton, NJ: Educational Testing Service.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Miller, K.F., Smith, C. M. Zhu, J., & Zhang, H. (1995). Preschool origins of cross-national differences in mathematical competence: The role of number-naming systems. *Psychological Science, 6*, 56-60.

Ministry of Education, Québec. (2002). *Evaluation Of Learning At The Preschool And Elementary Levels: Framework*, Gouvernement du Québec, Ministère de l'Éducation.

Molenaar, P.C.M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*, 201–218.

Molenaar, P.C.M., & Campbell, C.G. (2009). The new person-specific paradigm in psychology, *Current Directions in Psychological Science, 18*, 112-117.

Molenaar, P.C.M., Sinclair, K.O., Rovine, M.J., Ram, N., & Corneal, S.E. (2009). Analyzing developmental processes on a individual level using non-stationary time series modeling. *Developmental Psychology, 45*, 260–271.

National Assessment Program Literacy and Numeracy (NAPLAN). (2008*). Achievement in reading, writing, language conventions and numeracy.* Melbourne, Australia: Ministerial Council on Education, Employment, Training and Youth Affairs.

National Center for Education Statistics (2005). *A First Look at the Literacy of America's Adults in the 21st Century*. NCES Number: 2006470

National Curriculum Board. (2009). *The shape of the Australian curriculum: English.* Canberra, Australia: Commonwealth of Australia 2009. Retrieved from http://www.acara.edu.au.

National Curriculum Council (UK). (1991). Report on monitoring the implementation of the national curriculum core subjects 1989-1990. York, England: Author.

National Report on Schooling in Australia Preliminary Paper. (2007). *National Benchmark Results Reading, Writing and Numeracy Years 3, 5 and 7*. Melbourne, Australia: Ministerial Council on Education, Employment, Training and Youth Affairs.

Neilson. (1913). From Professor Neilson. *The English Leaflet, 104*, 5. Retrieved from http://www.archive.org/details/hillegasscale00newerich

Northwest Evaluation Association. (2002). *RIT scale norms*. Retrieved from http://www.nwea.org/

Northwest Evaluation Association. (2005). *2005 normative data: Monitoring growth in student achievement*. Retrieved from http://www.nwea.org/

Northwest Evaluation Association. (2005). *Normative data, RIT point monitoring, growth in student achievement.* Retrieved from http://www.nwea.org/

Olsen, L.W. (2003). *Essays on Georg Rasch and his contributions to statistics*. PhD thesis-part, Institute of Economics, University of Copenhagen. Retrieved from http://www.rasch.org/olsen.htm

Ouchi, W.G. (2009). The Secret of TSL: The Revolutionary Discovery That Raises School Performance. New York, NY: Simon & Schuster.

Oxford University Archives (n.d.). Retrieved from http://www.oua.ox.ac.uk/holdings/Local%20Examinations%20Delegacy%20LE.pdf

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J., & Kirby, J.R. (2005). Development of individual differences in reading: Results from longitudinal studies in English and Finnish. *Journal of Educational Psychology, 97*, 299–319.

Participation and Achievement Series:Nos 2-5. (1998, 1999). Broadsheet pamphlets reporting profile data for English, science, mathematics, health, physical education. Adelaide, Australia: Department of Education, Training and Employment.

Patz, R.J. (2007). *Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems.* Prepared for the Technical Issues in Large Scale Assessment

(TILSA) State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO). CTB/McGraw-Hill.

Pedulla, J.J., Airasian, & Madaus, G.F. (1980). Do Teacher Ratings and Standardized Test Results of Students Yield the Same Information? *American Educational Research Journal, 17*, 3, 303-307.

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* National Research Council. Washington, DC: National Academy Press.

Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments*. National Center for the Improvement of Educational Assessment. Retrieved from http://www.nciea.org/cgi-bin/pubspage.cgi?sortby=pub_date

Perry, N., & Meisels, S. (1996). *How accurate are teacher judgments of students' academic progress*? National Center for Educational Statistics, Working Paper 96-08. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=9608

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). *Scaling, norming, and equating.* In R.L. Linn, *Educational measurement* (3rd ed. pp. 221-262). New York: Macmillan.

Piper, K.R. (1997). *Riders in the chariot: Curriculum reform and the national interest: 1965-1995*. Melbourne, Australia: Australian Council for Educational Research.

Pollack, J.M., Atkins-Burnett, S., Najarian, M., & Rock, D.A. (2005). Early Childhood Longitudinal Study, Kindergarten class of 1998–99 (ECLS–K): Psychometric report for the fifth grade. (NCES 2006–036). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006036rev

Popham, W.J. (2003). *Crafting curricular aims for instructionally supportive assessment*. Paper prepared for American Association of School Administrators, the National Association of Elementary School Principals, the National Association of Secondary School Principals, the National Education Association, and the National Middle School Association.

Popham, W.J. (2007). The lowdown on learning progressions, *Educational Leadership*, *64*(7), 83-84.

Popham, W.J., & Husek, T.R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*, 1–9.

Popham, W.J., Cruse, K.I., Rankin, S.C., Sandifer P.D., & Williams P.L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66*, 628–634.

Qualifications and Curriculum Authority. (2008). *Assessment and reporting arrangements 2009*. Retrieved from http://www.naa.org.uk/naa_21551.aspx

Qualifications and Curriculum Authority. (2009a). *Evaluation of the assessing pupils' progress in year 1 pilot project 2007/8 Final report*, September 2008 QCA/09/4016. Retrieved from http://www.qcda.gov.uk/resources/publication.aspx?id=5079a52a-e09b-4001-be36-19e66fd13075

Qualifications and Curriculum Authority. (2009b). *Single level tests: Report of the first three test sessions: December 2007, June 2008 and December 2008*. QCA/09/4303. Retrieved from http://www.qcda.gov.uk/resources/ [link broken]

Qualifications and Curriculum Development Agency. (2010). *Single level tests arrangements*. Retrieved from http://testsandexams.qcda.gov.uk/

Quantile Framework for Mathematics. (2010). Metametrics. Retrieved from www.Quantiles.com

Quinlan, M., & Scharaschkin, A. (1999, September). National curriculum testing: problems and practicalities. Paper presented at the British Educational Research Association Annual Conference, Brighton. Cited by Tymms (2004) and Durant (2003).

Raudenbush S.W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, 52, 501–525.

Rao, C.R. (1958). Some statistical methods for comparison of growth curves. *Biometrics, 14*, 1-17.

Rogosa, D.R., and Willett, J.B. (1985). Understanding Correlates Of Change By Modeling Individual Differences In Growth. *Psychometrika*, *50*, 203-228.

Raudenbush, S.W. (2009). The Brown legacy and the O'Connor challenge: Transforming schools in the images of children's potential, *Educational Researcher, 38*, 169-180.

Ravitch, D. (2010). The death and life of the great American school system: How testing and choice are undermining education. New York, NY: Basic Books.

Reid, A. (1991). *Attainment levels: A critique*. Paper presented at a seminar, Underdale Campus University, Centre for Studies In Educational Leadership.

Reid, A. (1992a). Reshaping education: National agenda - national curriculum problems and possibilities, *SAIT Journal,* 24 June, 1992 p.13 and p.22.

Reid, A. (1992b). Accountability and education: The great profile debate. *Curriculum Perspectives*, *12*(1), 55-57.

Reid, A. (1995). Profiles: Real problems or real gains-from whose perspective? *Curriculum Perspectives*, *15*(3), 76-80.

Reid, A. (1999). The national education agenda and its curriculum effects. In B. Johnson & A. Reid (Eds.). *Contesting the curriculum*. Katoomba, Australia: Social Science Press.

Reid, A. (2005). Rethinking national curriculum collaboration: Towards an Australian curriculum. Canberra, Australia: DEST.

Rice, J.M. (1893). *The public-school system of the United States*. New York, NY: The Century Company. Retrieved from www.archive.org/details/ scientificmanag00ricegoog.

Rice, J.M. (1897). The futility of the spelling grind. *Forum*, 23, 163-172, 409-419. Republished in J.M. Rice (1913). *Scientific management in education*. New York, NY: Publishers Printing Company. Retrieved from www.archive.org/ details/scientificmanag00ricegoog.

Rice, J.M. (1902). Educational research: a test in arithmetic. *Forum*, 34, 281-297. Republished in J.M. Rice (1913). *Scientific management in education*. New York, NY: Publishers Printing Company. Available at www.archive.org/details/ scientificmanag00ricegoog.

Rice, J.M. (1913). *Scientific management in education*. New York, NY: Publishers Printing Company. Available at www.archive.org/details/scientificmanag00ricegoog.

Rogoff, B. (1990). Apprenticeship in thinking: Cognitive development in social context. New York, NY: Oxford University Press.

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom: teacher expectation and pupils' intellectual development. New York, NY: Holt, Rinehart, & Winston.

Rosenthal, R., & Rubin, D.B. (1978). Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences*, 3, 377–386.

Rothman, S. (1998). *Factors influencing assigned student achievement levels*. Paper presented at Australian Association for Research in Education. Retrieved from http://www.aare.edu.au/98pap/rot98337.htm

Rothman, S. (1999). *Factors influencing assigned student achievement levels II: Mathematics, The Arts, and Health and Physical Education*. Paper presented at Australian Association for Research in Education. Retrieved from http://www.aare.edu.au/99pap/rot99019.htm

Rowe, K.J., & Hill, P.W. (1996). Assessing, recording and reporting students' educational progress: The case for 'subject profiles'. *Assessment in Education, 3*, 309–352.

Rudner, L.M., & Boston, C. (2003). Data warehousing: Beyond disaggregation. *Educational Leadership, 60*(5), 62-65.

Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*(4), 219-262.

Ryan, J., & Williams, J. (2007). *Children's mathematics 4-15: Learning from errors and misconceptions*. Buckingham, England: Open University Press.

Sadler, D.R. (1986). *Subjectivity, objectivity and teachers' qualitative judgements.* Assessment Unit Discussion Paper 5. Brisbane, Australia: Board of Secondary School Studies.

Sadler, D.R. (1987). Specifying and Promulgating Achievement Standards, *Oxford Review of Education, 13*(2), 191-209.

Sadler, D.R. (1998). Formative assessment: revisiting the territory. *Assessment in Education*, *5*, 77-84.

Schulz, E.M., & Nicewander, W.A. (1997). Grade equivalent and IRT representations of growth. *Journal of Educational Measurement*, *34*(4), 315-331.

Shepard, L.A. (2000) *The Role of Classroom Assessment in Teaching and Learning*. CSE Technical Report 517. CRESST/University of Colorado at Boulder.

Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F., Snowden, J.B., Gordon, E., Gutierrez, C., & Pacheco, J. (2005). Chapter 8 in Darling-Hammond, L. and Bransford, J. (Eds) *Preparing teachers for a changing world: What teachers should know and be able to do*. San Francisco, CA: Jossey-Bass.

Sharpley, C.F., & Edgar, E. (1986). Teachers' ratings vs standardized tests: an empirical investigation of agreement between two indices of achievement. *Psychology in the Schools, 23*, 106–111.

Shinn, M.R. (Ed.) (1989). *Curriculum-based measurement: Assessing special children.* New York, NY: The Guilford Press.

Shute, V.J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153-189.

Singer, J.D., & Willett, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Slinde, J.A., & Linn, R.L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement, 15*, 23-35.

Smith, N.J. (2008). Next generation state data system: What is needed to support the next generation assessment and accountability systems. Data Quality Campaign. Retrieved from www.DataQualityCampaign.org

Sneyd-Kynnersley, E.M. (1913). HMI: Some passages in the life of one of H.M. inspectors of schools. London, England: MacMillan.

South Australian Curriculum, Standards and Accountability Framework. (2000). *Implementation plan 2001 to 2002*. Adelaide, Australia: Department of Education, Training and Employment.

Spady, W. (1993). *Outcomes-based education.* Workshop Report Number 5. Canberra, Australia: ACSA,.

Stake, R.E. (Ed). (1991). Using assessment policy to reform education:, Advances in program evaluation. Greenwich, CT: JAI Press.

Star Reading. (2005). *Star reading computer adaptive test and database technical manual*. Wisconsin Rapids, WI: Renaissance Learning, Inc.

Starch, D. (1916). *Educational Measurements*. New York, NY: The Macmillan Company.

Starch, D., & Elliott, E.C. (1912). Reliability of grading high school work in English, School Review, 20, 442-457.

State Educational Technology Directors Association (SETDA). (2008). *Technology-Based Assessments Improve Teaching and Learning*. Retrieved from http://www.setda.org/web/guest/2020

Statistical First Release 29/1999 (1999). *National Curriculum Assessments of 7, 11 and 14 year olds by Local Education Authority 1999*. Department for Children, Schools and Families. Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/index.shtml

Statistical First Release 43/2000 (2000). *National Curriculum Assessments of 7, 11 and 14 year olds by Local Education Authority 2000*. Department for Children, Schools and Families. Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/index.shtml

Statistical First Release 37/2001 (2001). *National Curriculum Assessments for 7, 11 and 14 year olds in England 2001.* Department for Children, Schools and Families. Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/index.shtml

Statistical First Release 21/2002 (2002). *National Curriculum assessments of 7 and 11 year olds in England 2002 (Provisional)*. Department for Children, Schools and Families. Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/index.shtml

Statistical First Release 20/2003 (2003). *National Curriculum Assessments of 7, 11 and 14 year olds in England, 2003 (Provisional).* Department for Children, Schools and Families. Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/index.shtml

Statistical First Release 30/2004 (2004). *National Curriculum Assessments of 11 year olds in England, 2004 (Provisional).* Department for Children, Schools and Families. Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/index.shtml

Statistical First Release 31/2005 (2005). *National Curriculum Assessments of 11 year olds in England, 2005 (Provisional)*. Department for Children, Schools and Families. Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/index.shtml

Statistical First Release 31/2006 (2006). *National Curriculum Assessments at Key Stage 2 in England, 2006 (Provisional).* Department for Children, Schools and Families. Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/index.shtml

Statistical First Release 24/2007 (2007). *National Curriculum Assessments at Key Stage 2 in England, 2007 (Provisional).* Department for Children, Schools and Families. Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/index.shtml

Statistical First Release 19/2008 (2008). *National Curriculum Assessments at Key Stage 2 in England, 2008 (Provisional).* Department for Children, Schools and Families. Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/index.shtml

Statistical First Release 20/2008. (2008). *National curriculum assessments at Key Stage 3 in England, 2008 (Provisional)*. Department for Children, Schools and Families. Retrieved from http://www.dcsf.gov.uk/rsgateway/DB/SFR/s000836/index.shtml

Statistical First Release 21/2008. (2008*). National curriculum assessments at Key Stage 1 in England, 2008.* Department for Children, Schools and Families. Retrieved from http://www.dcsf.gov.uk/rsgateway/DB/SFR/s000806/index.shtml

Statistical First Release 06/2009 (2009). *National curriculum assessments at Key Stage 2 in England, 2008 (Revised)*. Department for Children, Schools and Families. Retrieved from http://www.education.gov.uk/rsgateway/DB/SFR/index.shtml

Statistical First Release 19/2009. (2009). *National curriculum assessments at Key Stage 2 in England, 2009*. Department for Children, Schools and Families. Retrieved from http://www.dcsf.gov.uk/rsgateway/DB/SFR/s000865/index.shtml

Stehn, J. (1997). Renewing the curriculum cycle in South Australia. In J. Lokan (Ed.). (1997). *Describing Learning; Implementation of Curriculum Profiles in Australian Schools 1986-1996*. (pp. 166-194). Melbourne, Australia: Australian Council for Educational Research.

Stenner, A.J., & Stone, M.H. (2004, May), 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. Retrieved from Metametrics website www.lexiles.com

Stenner, A.J., Burdick, H., Sanford, E.E., & Burdick, D.S. (2007). The Lexile framework for reading technical report. Retrieved from Metametrics website www.lexiles.com.

Stenner, J. (1996). *Measuring reading comprehension with the Lexile framework.* Paper presented at the Fourth North American Conference on Adolescent/Adult Literacy. Washington, D.C. February 1996.

Stiggins, R. (2002). Leadership for Excellence in Assessment, A Powerful New school District Planning Guide. Portland, Oregon: Assessment Training Institute.

Stiggins, R.J. (2008). Assessment Manifesto: A Call for the Development of Balanced Assessment Systems. Princeton, N.J.: Educational Testing Service.

Stobart, G. (2001). The validity of national curriculum assessment. *British Journal of Educational Studies*, *49*(1), 26–39.

Stone, C.W. (1908) Arithmetical Abilities, and Some Factors Determining Them. *Teacher's College Series No 19*, Columbia University.

Strøm, B. (2004). *Student achievement and birthday effects. Mimeo, Norwegian University of Science and Technology.* Retrieved from http://www.hks.harvard.edu/pepg/PDF/events/Munich/PEPG-04-24Strom.pdf

Suppes, P., Fletcher, J.D., & Zanotti, M. (1976). Models of individual trajectories in computer-assisted instructions for deaf students. *Journal of Educational Psychology, 68*, 117-127.

Test Statistics. (2007). *Key stage 2 – English, mathematics and science test statistics.* Retrieved from http://www.naa.org.uk/naa_19225.aspx

National Council of Teachers of English (1917). Proceedings of the sixth annual meeting, New York City, November 30 and December 1-2, 1916. *The English Journal*, *6*(1), 40-68.

Thomas, C.S. (1913). The Hillegas Scale, *The English Leaflet, 104*, 1-4. Retrieved from http://www.archive.org/details/hillegasscale00newerich

Thomson, P. (1999). How doing justice got boxed in: a cautionary curriculum tale for policy activists. In B. Johnson & A. Reid (Eds.). *Contesting the curriculum*. Katoomba, Australia: Social Science Press.

Thorndike, E.L. (1910). Handwriting. *Teachers College Record, 11*(2), 1-81. Retrieved from http://www.tcrecord.org

Thorndike, E.L. (1911). A scale for measuring the merit of English writing. *Science*, *33*(859), 935-938.

Thorndike, E.L. (1912). The measurement of educational products. *The School Review*, *20*(5), 289-299.

Thorndike, E.L. (1913). Notes on the Significance and Use of the Hillegas Scale for Measuring the Quality of English Composition. *The English Journal, 2*(9), 551-561.

Thorndike, E.L. (1914a). Educational psychology volume 3: Mental work and fatigue and individual differences and their causes. New York, NY: Teachers College, Columbia University.

Thorndike, E.L. (1914b). The measurement of ability in reading: Preliminary scales and tests. *Teachers College Record*, *15*(4), 1-71. Retrieved from http://www.tcrecord.org

Thorndike, E.L. (1915). An improved scale for measuring ability in reading. *Teachers College Record*, *16*(1), 31-53. Retrieved from http://www.tcrecord.org

Thorndike, E.L. (1916a). *An introduction to the theory of mental and social measurements*. (2nd Ed.). New York, NY: Teachers College, Columbia University.

Thorndike, E.L. (1916b). Measurement of achievement in reading: Word knowledge. *Teachers College Record*, *17*(5), 430-454. Retrieved from http://www.tcrecord.org

Thorndike, R. L. (1971). *Educational measurement*. (2nd ed.) Washington, DC: American Council on Education.

Thorndike, R.M., & Lohman, D. (1990). *A century of ability testing*. Chicago, IL: Riverside.

Thurber. (1913). From Mr. Thurber. *The English Leaflet, 104*, 5-6. Retrieved from http://www.archive.org/details/hillegasscale00newerich

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433-451.

Thurstone, L.L. (1928). Scale construction with weighted observations. *Journal of Educational Psychology, 19*, 441-453.

Timperley, H. (2009, August). *Using assessment data for improving teaching practice*. Paper presented to the Australian Council for Educational Research Assessment and Student Learning Conference, Perth.

Tourangeau, K., Nord, C., Lê T., Pollack, J.M., and Atkins-Burnett, S. (2006). Early Childhood Longitudinal Study, Kindergarten class of 1998–99 (ECLS-K), Combined user's manual for the ECLS-K Fifth-grade data files and electronic codebooks (NCES 2006–032). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Trabue, M.R. (1916). Trabue's dissertation on completion tests. *Teachers College Record*, *17*(1), 85-91. Retrieved from http://www.tcrecord.org

Trabue, M.R. (1917) Supplementing the Hillegas Scale. *Teachers College Record, 18*(1), 51-84. Retrieved from http://www.tcrecord.org

Training and Development Agency for Schools. (2008). *Professional Standards for Qualified Teacher and Requirements for Initial Teacher Training.* Retrieved from http://www.tda.gov.uk/teachers/professionalstandards/downloads.aspx

Triga, A. (2004). An analysis of teachers' rating scales as sources of evidence for a standardised Greek reading test. *Journal of Research in Reading*, 27, 311–320.

Tymms , P., Merrell, C., & Jones, P.(2003, October). *The problems of creating a one-entry assessment across culture and languages.* Paper presented to the 29th International Association for Educational Assessment (IAEA) Annual Conference.

Tymms, P. (2004). Are standards rising in English primary schools? *British Educational Research Journal*, *30*, 477-494.

Tymms, P., & Merrell, C. (2007). *Standards and quality in English primary schools over time: the national evidence.* Primary Review Research Survey 4/1. Cambridge, England: University of Cambridge Faculty of Education.

Tymms, P., & Wylde, M. (2003, April). *Baseline assessment and monitoring in primary schools*. Paper presented at the Symposium Connectable Processes in Elementary and Primary Section. Bamberg. Retrieved from www.cemcentre.org/publications/ downloads/germansymposium.doc

Tymms, P., Merrell, C., & Jones, P. (2004). Using baseline assessment data to make international comparisons, *British Educational Research Journal*, *30*, 673-689.

Victorian Auditor-General. (2009). *Literacy and numeracy achievement*. Melbourne, Australia: Victorian Government Printer.

Victorian Curriculum and Assessment Authority. (2004). *Victorian curriculum reform 2004 consultation paper*. Retrieved from www.vcaa.vic.edu.au

Victorian Curriculum and Assessment Authority. (2006a). *VELS Standards and Progression Points: English*. Retrieved from http://vels.vcaa.vic.edu.au

Victorian Curriculum and Assessment Authority. (2006b). *VELS Standards and Progression Points: Mathematics*. Retrieved from http://vels.vcaa.vic.edu.au

Victorian Curriculum and Assessment Authority. (2009). *National Assessment Program Literacy and Numeracy (NAPLAN) reference guide – Analysing NAPLAN data*. Retrieved from www.vcaa.vic.edu.au/prep10/naplan/schools/ analysingnaplandata.pdf

Victorian Institute of Teaching. (2009). *Standards For Graduating Teachers*. Retrieved from http://www.vit.vic.edu.au/content.asp?Document_ID=23

von Davier, A.A., Holland, P.W., & Thayer, D.T. (Eds.). (2004). *The kernel method of test equating*. New York, NY: Springer-Verlag.

Vygotsky, L.S. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

Walston, J., Rathbun, A., & Germino Hausken, E. (2008). *Eighth grade: First findings from the final round of the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99*. (NCES 2008-088). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement, 69*, 760-777.

Wang, S., Jiao, H., Brooks, T., & Young, M.J. (2004). *Construct equivalence between customized and original Stanford Achievement Reading Comprehension Tests* (10th ed.) Research Report. San Antonio, TX: Harcourt Assessment. Cited in Wang, S. & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12

large-scale reading assessment. *Educational and Psychological Measurement, 69*, 760-777.

Wasik, B.H., & Loven, M.D. (1980). Classroom observational data: Sources of inaccuracy and proposed solutions. *Behavioral Assessment, 2*, 211-227.

Wiliam, D. (1994). *Towards a philosophy for educational assessment.* Revised version of paper originally presented to the British Educational Research Association's 20th annual conference in Oxford in 1994. Retrieved from http://www.kcl.ac.uk//depsta/education/hpages/dwliam.html

Wiliam, D. (1998, September). Enculturating learners into communities of practice: Raising achievement through classroom assessment. Paper presented at European Conference on Educational Research, Ljubljana, Slovenia. Cited by Forster (2009).

Wiliam, D. (1999). Standards: Annex 4 in Weighing the baby: the report of the independent scrutiny panel on the 1999 Key Stage 2 national curriculum tests in English and mathematics. London, Department for Employment and Education. Cited by Stobart, G. (2001). The validity of national curriculum assessment. British Journal of Educational Studies, 49, 26–39.

Wiliam, D. (2010). What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment. *Review of Research in Education*, *34*, 254-284.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C.A. Dwyer.(Ed.) *The future of assessment: shaping teaching and learning*. Mahwah, NJ: Erlbaum.

Willett, J.B., & Sayer, A.G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116*, 363-381.

Williams, D., Johnson, B., Peters, J., & Cormack, P. (1999). Assessment: from standardised to authentic approaches. In B. Johnson & A. Reid (Eds.). *Contesting the curriculum*. Katoomba, Australia: Social Science Press.

Williams, J., Wo, L., & Lewis, S. (2007). Mathematics progression 5-14: Plateau, curriculum/age and test year effects. *Research in Mathematics Education, 9*, 127-142.

Williamson, G.L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal analyses of academic achievement. *Journal of Educational Measurement*, *28*, 61-76.

Williamson, G.L. (2006). *What is expected growth?* A white paper from The Lexile Framework for Reading. Retrieved from Metametrics website www.lexiles.com

Wilson, B. (1993). *National curriculum and national profiles: Development, implementation and use.* Incorporated Association of Registered Teachers of Victoria Seminar Series No 23. Melbourne, Australia.

Wilson, B. (1994). Profiles meet post structuralism. *Curriculum Perspectives*, *14*(2), 20-21.

Wilson, M. (2004). Assessment, Accountability and the Classroom: A Community of Judgment. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability.* 103rd Yearbook of the National Society for the Study of Education. (pp. 1-19). Chicago, IL: University of Chicago Press.

Wilson, M. (Ed.). (2004). *Towards coherence between classroom assessment and accountability*. 103rd Yearbook of the National Society for the Study of Education. Chicago, IL: University of Chicago Press.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*, 181-208.

Withers, G. (1999). *Supporting the implementation of statements and profiles in South Australian schools*: Report of a survey conducted in November – December 1998. Melbourne, Australia: Australian Council for Educational Research.

Wright, B.D. (1968). Sample-free test calibration and person measurement. Proceedings 1967 Invitational Conference on Testing Princeton: Educational Testing Service, 85-101. Cited in Wright (1977).

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97-116.

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago, IL: MESA Press.

Wright, B.D., & Stone, M.H. (1999). *Measurement essentials. (*2nd ed.). Wilmington, DE: Wide Range Inc. Chapter retrieved from www.rasch.org/measess/met-9.pdf

Wright, B.D., & Stone, M.H. (2004). *Making measures*. Chicago, IL: The Phaneron Press.

Wright, B.D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *16*(4), 33-45.

Wright, B.D. (2001). Counts or measures? Which communicate best? *Rasch Measurement Transactions, 14*(4), 784.

Wright, D. & Wiese, M.J. (1988). Teacher judgement in student evaluation: A comparison of grading methods. *Journal of Educational Research, 82*, 10-14.

Wyatt, H.G. (1917). *Methods of school inspection in England.* Occasional Reports No. 7. , India: Bureau of Education. Retrieved from http://www.archive.org/details/methodsofschooli00wyatrich

Yeh, S.S. (2006). Raising student achievement through rapid assessment and test reform. New York, NY: Teachers College Press.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125-145.

Yen, W.M. 1985. Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50*(4), 399–410.

Yen, W.M. 1986. The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*(4), 299–325.

Appendix 1 Letter of approval for data access

Department of Education
and Children's Services

Office of People and Culture

Education Centre
31 Flinders Street
Adelaide 5000
South Australia
GPO Box 1152
Adelaide 5001

DECS CS/04/4934.6

15 February 2005

Mr Martin Caust
PO Box 6285
James Cook University – Education Faculty
O'CONNOR ACT 2602

Dear Mr Caust

Thank you for your letter requesting approval for your project *'Re-analysis of 1997 and 1998 Profiles collection, and, in addition, correlation of profile assessment'*.

Your project has been reviewed by a senior DECS consultant with respect to protection from harm, informed consent, confidentiality and suitability of arrangements. Subsequently, I am pleased to advise you that after careful consideration your project has been **approved**.

If changes are to be made to your proposal, please supply the department with an electronic copy of the changes made as well as an electronic copy of the final report, which will be circulated to interested staff and then made available to DECS educators for future reference.

I wish you well with your project.

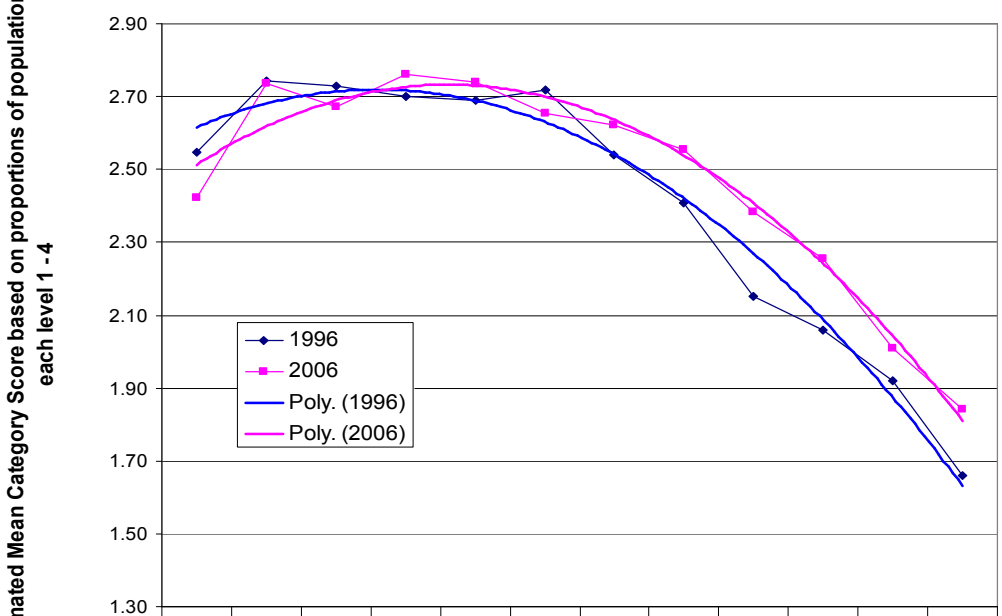
Lexie Mincham
MANAGER, NETWORKED LEARNING COMMUNITY

Government of South Australia

Appendix 2 Adult literacy trends with age

Figure A2.1 is based on 2006 ABS data. For both 1996 and 2006 average prose reading skill levels increase from Age 15-19 up to about 30-35 and taper off after that. This implies an increase in skill level across the whole age cohort even after leaving school. It is possible this is due to ongoing enrolment in post-secondary education.

Figure A2.1 Adult Literacy Australia- Trend in Prose Scale by Age category

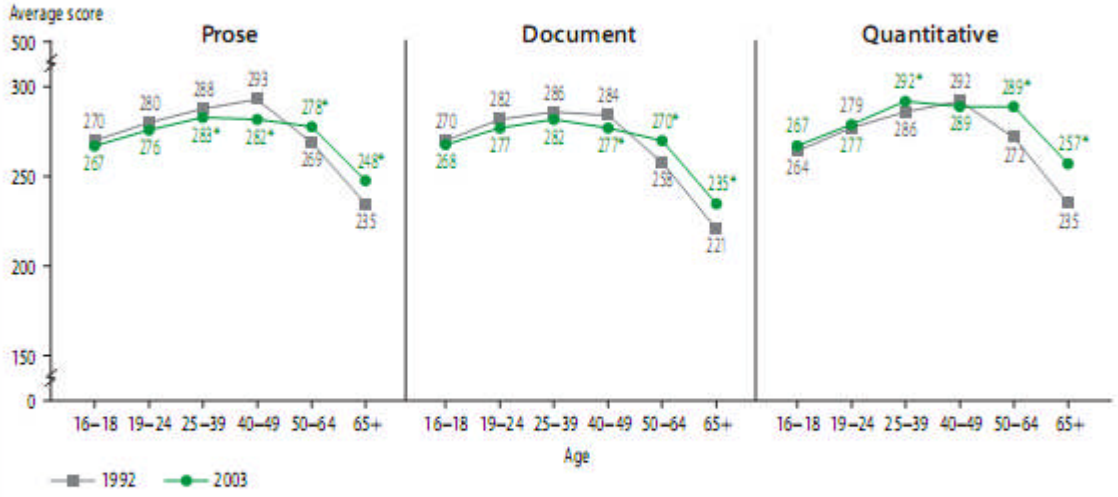


	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74
1996	2.55	2.74	2.73	2.70	2.69	2.72	2.54	2.41	2.15	2.06	1.92	1.66
2006	2.42	2.74	2.67	2.76	2.74	2.65	2.62	2.56	2.38	2.26	2.01	1.84

Source: ABS 42280DO001 Adult Literacy and Life Skills Survey, Summary Results, 1996 and 2006

US reports of adult literacy provide direct scale score means for each age group. These are shown in Figure 11 below taken from the National Center for Education Statistics (2005). *A* First Look at the Literacy of America's Adults in the 21st Century. NCES Number: 2006470. The pattern by mean scale score by age category is similar to the Australian data. In both 1992 and 2003 the means scores increase until about age 40-49. Both patterns suggest increasing competency in literacy and quantitative skills with age, plateauing in the early to mid 30s.

Figure 11. Average prose, document, and quantitative literacy scores of adults, by age: 1992 and 2003



*Significantly different from 1992.

NOTE: Adults are defined as people 16 years of age and older living in households or prisons. Adults who could not be interviewed due to language spoken or cognitive or mental disabilities (3 percent in 2003 and 4 percent in 1992) are excluded from this figure.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 1992 National Adult Literacy Survey and 2003 National Assessment of Adult Literacy.

Source: National Center for Education Statistics (2005) A First Look at the Literacy of America's Adults in the 21st Century. NCES Number: 2006470.

Appendix 3 Adequacy of the Key Stage Test Assessments

There is some concern about the adequacy of the Key Stage test assessments. The use of the Key Stage tests as a reference for teacher assessment assumes that the tests are of adequate validity and reliability. Stobard (2001) considers the validity of the Key Stage tests and concludes that

> Even though the test development process is exemplary (Wiliam, 1999), the validity of their use as the key measure of year-on-year changes in national standards is questionable. …. (some) fluctuations may best be explained by problems with test equating rather than dramatic changes in standards in particular subjects. The consistency of Teacher assessment scores supports this interpretation.(Stobard, 2001, p. 32)

Stobard cites the official data on tests and teacher assessments for Key Stage Three over the period 1997 to 2001 (Stobard, 2001, p.33, Table 1) which show a pattern of approximate comparability. Thesis Figure 4.1 shows an equivalent trend for Key Stage 2. The percentages of students reported at level 4 in Figure 4.1 or higher from the two assessment sources differ from each other by as much as 6 percentage points (science in 2000) but most often differ by 3 points or less. There is greater variability in the test results around the general trend than the teacher assessments, although both follow approximately similar trends. Stobard argues that the greater stability of the teacher assessments is an indicator of the lesser validity of the test results (p.32).

Uncertainty about the consistency of the KS tests over time is shared by others. Tymms (2004) raises a range of concerns about the adequacy of the Key Stage tests as measures of performance of students over time. He queries whether cross-sectional cohorts of students are moving up the levels scales in English and mathematics over a sequence of years at the rate indicated by the official figures. He identifies two periods of change in student scores: 1995 to 2000, where scores rose rapidly and from 2001 to 2004 where scores appeared to hardly change. Student time series scores are represented at Key Stage 2 by the rather gross measure of 'percentage at or above level 4' in the Key Stage 2 tests. This indicator moves from being about average (of the order of 50% of students above and below the position in 1995) to the indicator showing over 70% of students above the position by 2003. Based on eleven data sources Tymms is concerned that the statutory tests might overstate the improvement from 1995 to 2000, and may have underestimated the rise from 2001 to 2004.

Part of the reason for the concern is the process for setting cut scores each year. Tymms citing Quinlan & Scharaschkin (1999) summarises the four processes applied as; "marker opinion, professional scrutiny of the test papers (Angoff techniques), earlier use of the live test and the employment of an anchor test" (Tymms, 2004, p. 489).

Two difficulties are identified by Tymms. The first is that the cut scores were required to correspond to a 'mark', that is an integer. He explains that "a change of one mark in 1996 would have made about a 1.4% difference in the proportion of students being awarded level 4 or above" (Tymms, 2004, p. 484). The effect in 2000 is estimated to be 1.8%.

The second difficulty is that prior to 2001 standards were only equated from one year to the previous year. Tymms argues that this allowed drift in relative standards due to the compounding of error over time. Post 2000 the equating process applied over multiple years, and "may well be the reason for the abrupt changes between Phases 1 and 2" (p. 490). (Phase 1 is the period 1995-2000; Phase 2 is the period 200-2004). Tymms concludes that the use of the tests in the period 1995 to 2004 for "monitoring standards over time…has failed for a number of reasons" (p. 492). As a result of this analysis some caution is needed in accepting the quality of the test data, especially when it is compared with, or used as a standard for,

teacher assessment. Unlike Stobard, Tymms has not used the teacher assessments as a possible guide to the 'real' change over the period.

Reference to the regular test statistics (Test Statistics, 2007) published on the QCA/NAA website provide an additional insight into the test development process. Tests in a given subject are developed by individual contractors (NFER, Edexcel as examples) without necessarily continuity of contract from one year to the next. The publicly reported analyses are very simple (Cronbach's alpha, percentages correct for each item in the trial samples, score ranges for mapping to levels i.e. 'level threshold tables'). The statistics summaries indicate that whatever analyses are made in test development, the allocation to a level is based on test marks as reported by Tymms. No item maps or indication of the spread of item difficulties for those items around the level boundaries are given. Individual student reports to parents are level only, not scores (*End of key stage 2 pupil results proforma*, 2008). The marked scripts are returned to schools, so schools are informed of 'marks', but the computer file/report provided with them indicates only a level in each subject (*Assessment and reporting arrangements* (ARA) 2009). Head teachers must report the result to parents 'within 15 school days of the head teacher receiving it' (ARA, 2009, p. 80). From the evidence of Tymms, Quinlan & Scharaschkin, Stobard and the QCA publications, the Key Stage tests and teachers assessments are reported to schools on a very broad scale. The subsequent summaries at Local Authority and national level used in times series reports are even broader, concerned mainly with the percentage of students that are at or above a particular level, depending on the Key Stage. Given the process, a general impact of measurement error is expected around the threshold for the appropriate level in each Key Stage.

Based on the evidence above, using the test data as the 'assumed' best possible independent estimate of a student's developmental position on the levels scale is problematic. Mismatch of teacher and test data would not necessarily indicate any inadequacy in the teacher assessment but may reflect some general looseness in the test data, even given the broadness of the level scale. However being aware of the patterns of the two assessment processes over time and the persistence of these patterns by subject, provides some hints as to the relationship.

Appendix 4 Scale changes CSF to VELS in Victoria

The Victorian Auditor-General's Report 2009 shows reduced statewide means at Year levels 3, 5, 7, & 9 from 2006 using teacher-assessed levels. The reduced means may be an effect of adding one more sub-level category to a level scale.

Aggregate data at a state level is reported in Appendix C (p. 69-70) of the Auditor General's report. The most recent years (2006 & 2007) show a drop for reading and number scores as reported by teachers of 0.05 to 0.1 of a level relative to 2005, at the point where the Victorian Essential Learning Standards (VELS) were introduced. The downward shift in teacher assessed state averages was consistent at year levels 3, 5, 7 and 9. The Auditor–General explains the drop thus:

> The lower achievement recorded from 2006 onwards most likely reflects the impact of the change in curriculum from the CSF to the VELS, which introduced higher standards of learning for students and a new curriculum and assessment system for teachers. (Victorian Auditor-General, 2009, p. 69)

Given the strong similarity of the VELS levels to the CSF levels in the main learning areas, that is virtually unchanged (Gough, 2006), it is unlikely that the discontinuity from 2006 onwards is due to a change in the descriptions and standards of the levels. The author believes there is an alternative explanation, given that the test measure is unchanged, and there is no major drop in performance in the test in 2006, the year of the introduction of the VELS.

The benchmark standard ('expected level') for teacher judgements is also reported as unchanged for 2006 or 2007 by the Auditor–General (p. 69), implying the Department of Education and Early Childhood Development (DEECD) has not assumed the standard has changed.

Thus another possible explanation for the consistent drop in teacher-assessed means is the change in the scale used in the recording of teacher judgement assessments, originally a three-category scale. Individual teacher assessments of students were recorded by teachers in school records as a level and then one of three zones within the level. The distributions and means, as reported back to schools as benchmarks (Department of Education, Victoria, 1997, 1998, 1999; Department of Education and Training, Victoria, 2002; Department of Education and Early Childhood Development, Victoria, 2003, 2006), are assumed to use a numerical conversion something like 4.17, 4.5, 4.83 as equally spaced values within a level (and between levels) to calculate means. The actual translation values used are not found in the documentation.

In 2006 there was a transition to a four-point scale recorded as 4.0, 4.25, 4.5, 4.75 as an example for level 4 (VELS Standards and Progression Points: Mathematics, 2006). This expansion of the scale introduces one new point, X.0, at the lower end of the scale for each level along with placing category 'centres' at slightly different points on the framework scale. The state mean teacher judgement data are likely to average to a lower value using the four scale points relative to the original three (based on author simulation 'experiments'). Student frequencies, originally spread over three categories, are spread over four categories, with the new category at the lower end of the scale within each level. At each point within each level there is likely to be a tendency to report slightly fewer cases, through the redistribution of the cases downwards over the four points[34]. In simulated data developed by the author, 7000

[34] Since 1998 teacher assessments have been collected electronically from schools (Department of Education Victoria, 1999). Implied in the description of the process is that these data are collected at

cases spread from 0 to 4.0 at values to two decimal places, had a mean of 2.209 when centred on three categories per level and a mean of 2.076 over four categories, a reduction of 0.13 of level. The reduction in the actual data is dependent upon the frequency of the cases across the full scale. For simplicity the reduction effect is assumed to be about 0.1 of a level.

The Auditor –General explains that the number of progression steps has increased from 2 to 3 (in quotation below), rather than from 3 to 4. This may turn on what is understood by 'progression steps'. Within a level a student can progress to two zones beyond the initial starting zone before achieving all the criteria for a level. A simpler description would use 3 positions expanded to 4 positions. The Auditor-General describes the situation thus:

> The assessment scale used by teachers for reporting includes progression steps between the standards to describe the incremental improvements students make in reaching each standard. The standards comprise Levels one to six, with three progression points at 0.25 VELS/CSF levels intervals between the standards. Since 1998 the scale has remained the same between curriculum changes, but with the introduction of the VELS in 2006 more advanced skills and knowledge were expected of students achieving the standards. The number of progression steps also increased, from two for the CSF, to three for the VELS. (Victorian Auditor-General, 2009, p. 17)

This example illustrates a consequence of varying the number of categories available on a scale. When making assessments on notionally similar scales, using different unit 'densities' on these scales will influence the comparison.

---

an individual student level. The Auditor-Generals' Report makes clear that up to late 2008 the department did not have the capacity to connect test and teacher assessments at the individual student level. A unique student number system to be introduced in 2009 will make this possible (Victorian Auditor-General's Report, 2009, p. 11).

Appendix 5 NAPLAN data and model.

The National Assessment Program—Literacy and Numeracy (NAPLAN) and the relationship of mean learning status of cohorts with Year level and age

Preliminary comment.

Through out this appendix (and in Appendices 6 and 7) the *CurveExpert* (Hyams, 2001) software is used to fit a model developed from the Gompertz expression (Gompertz, 1825). The curve fitting is a pragmatic process to idealise the trajectories. Alternative curves can serve this purpose. A quadratic curve can provide an approximately equivalent solution (as can some other models applied in biological research included in the *CurveExpert* software). The Gompertz expression is selected for two reasons. The less important is that it pays homage to Courtis who originally proposed it as a family of curves to model growth in learning over time (see footnote 7 in chapter 2). The more important reason is pragmatic. The smooth curve derived for learning improvement with age fits the data points well. However, in addition, the inflection points for the various curves that can be fitted to the curve of the means or to the trajectories of +/- 1.96 SD, offer further mathematical modelling of slow and faster moving cohorts (relative to the mean cohort) in ways that provide hypotheses for further research into learning growth rates with age. In the process models for the trajectories of SD at particular ages are also predicted. However, for the basic modelling of the means of learning status with age a quadratic function works adequately.

NAPLAN

The first Australian National Assessment Program—Literacy and Numeracy (NAPLAN) tests were conducted in May 2008 for all Years 3, 5, 7 and 9 students in government and non-government schools (National Assessment Program Literacy and Numeracy (NAPLAN) Full Report, 2008). This publication has been timely and helpful in the refinement of the understanding of the general trend in test performance with increasing Year levels and age for this thesis. Prior to the NAPLAN report, information from some US states and from some US test norming programs, had been the best sources for explaining (and justifying for extrapolation and interpolation) some of the 'dynamics' of learning in English language and mathematics over an extended period of schooling. While the NAPLAN data are cross-sectional, based on Hilton & Patrick (1970) the trends can be considered as approximately similar to the longitudinal situation and indicative of the likely trends that applied in the South Australian data of 1997 and 1998.

The 2008 collection was the first occasion that students in Australia had been assessed at these Year levels with a set of common and linked tests and an underlying vertical scale. From 1999 to 2007 the National Report on Schooling published data, commencing with reading in 1999 at Years 3 and 5 only. By 2007 the National Report included reading, writing and numeracy at Years 3, 5 and 7. However students sat different tests in each State and Territory and while these were referenced to national benchmarks there were no reported vertical scales. Results were reported as percentages meeting national benchmark criteria (National Report on Schooling in Australia, 2007).

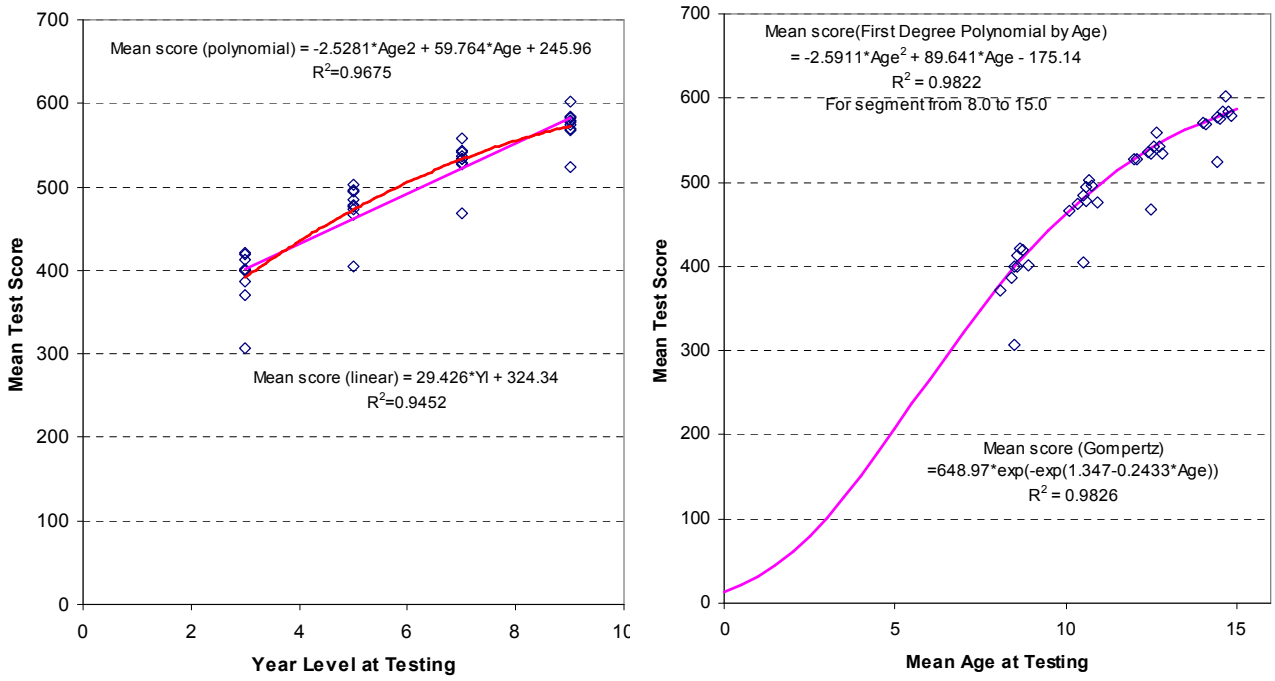
For the vertical scale, skills and understandings assessed in each domain (reading, writing, spelling, grammar and punctuation, and numeracy) from Year 3 through to Year 9 are mapped onto achievement scales with scores that range from 0 to 1000. By locating all students on a single national scale for each domain, the scales provide a clearer understanding of the general trends of learning with time/age/Year level. The 'reporting scales are constructed so that any given scaled score represents the same level of achievement over time. For example,

a score of 700 in Reading will have the same meaning in 2010 as it has in 2008' (NAPLAN Full Report, 2008, p. 3).

Modelling the learning trajectory

The 2008 data for each Year level are reported in Figure A5.1 in the left panel. Each point represents one state or territory system with government and non-government schools combined. In the right panel, the same data are graphed by average age at testing rather than by Year level. (The data could also be presented by elapsed years of schooling rather than age.) The four low outlier points in each panel are the points for the Northern Territory. These points are omitted for model estimation using 'curve-fitting' software (Hyams, 2001, Curve Expert). The vertical axis is plotted using the original 1000 point scale for the test collection scale. As will be indicated later in this appendix, the assumption about where the test scale origin at time 0 (birth) occurs has an influence on the trajectory of the curve, particularly in the lower ages and also on the model determined final asymptote. Varying the notional test scale score at 0 time also has an influence on the location of the inflection point in sigmoid models. For convenience, the original data are left unchanged, that is the original scale is preserved. The fitted model in this cases passes through Test score =13 at age=0. The Gompertz model assumes that the scale has a value on the vertical axis close to zero at time zero.

Figure A5.1 Plot of NAPLAN System Averages for Reading-2008



In the left panel the effect of centreing points on Year level on the time dimension eliminates the spread of the individual points on the time axis. The Coefficients of Determination (R^2) are shown as one indicator of model fit. For linear fit (Northern Territory omitted) R^2 is 0.9452. The R^2 for a first-degree polynomial is 0.9675.

In the right panel the data points are spread along the time axis at the average age at testing. Two curves are fitted, which at the scale of the graph, appear to be visually identical. The first-degree polynomial fitted from age 8 to 15 (R^2 = 0.9822 –NT omitted), follows approximately the same trajectory as the Gompertz curve (R^2=0.9826 – NT omitted) for the same segment. The Gompertz curve is plotted from age 0 to age 15.

A variety of solutions for a mathematical description of a smoothed curve to model the data points are possible. As described in Chapter 2 and in Chapter 5, the Gompertz curve provides a good fit for the points both statistically and conceptually and brings with it some 'testable' spin-offs related to rates of learning at points on the curve. The relationship between rate of leaning and the dispersion of the data (SD) near the inflection point of the rate of learning for a cohort/groups at specific ages can be explored. As a result the general first order Gompertz model is used to model the trajectory of the mean learning status of various cohorts in the subsequent sections of this appendix.

The Gompertz curve for the model in Figure A5.1 has the form:

Mean Test score = 648.97*exp(-exp(1.347-0.2433*Age)), where 648.97 is the asymptote for the curve.

What is clear from Figure A5.1 is that the general trend for the cross-sectional mean cohort scores by age and Year level for reading is non-linear. The left panel indicates a better fit for a quadratic function than a linear function. In both panels the trajectory can be modelled by a Gompertz model, a quadratic model and a number of other unreported possibilities (Morgan Mercer Flodin (MMF) as one example). The right panel shows a relationship of the mean test scores within systems to the age at testing. From this example there is supporting evidence for allowing the use of a non-linear model for interpolating and extrapolating cohort data discussed in more detail in Chapter 6. A simplification of the model using national norms rather than using each of the states as data points produces an equivalent model. However the model based on state data points highlights an apparent anomaly in state comparisons when data are centred on Year levels. These state comparisons are discussed briefly at the end of the appendix..

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

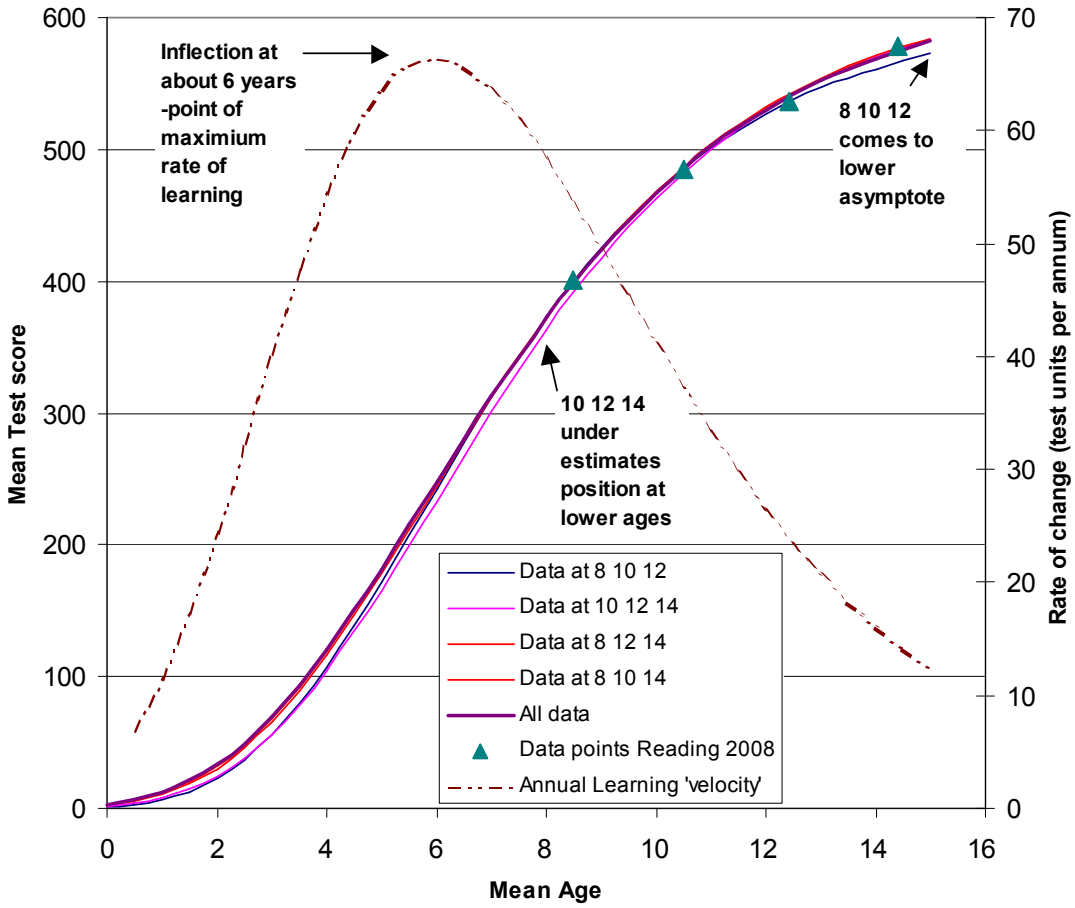
Based on the approximate fit of the Gompertz relation to the individual state and territory data, a model based on the national means can be developed as a general indicator of the trajectory for group means with age in reading and numeracy. On this basis the model is based on almost all students in Australia in Years 3, 5, 7, 9 in 2008. The relationships identified in this model development provide some general insights to be applied in the model for SA data for 1997 and 1998.

The impact of missing data points on curve parameters

The four means for each age cohort enable the impact on the parameters estimated for the curve when data points are missing to be explored along with the impact of these missing points on the trajectory fitted. This exploration will influence approaches adapted in Chapter 6 for extrapolation, where only two real points (Year 3 and Year 5) are available in the calendar years of the data reported, although proxies for Year 7 and Year 9 can be considered, based on more recent data.

Figure A5.2 is developed from the national reading data. The graph provides a comparison of the application of the Gompertz model when varying data points from among the set of available data points are used to estimate the curve. The models are fitted to the national mean for reading for each Year level cohort, plotted at its average age. Age cohorts are identified as 8, 10, 12 and 14 although the actual points are plotted at 8.5, 10.5, 12.4 and 14.4.

Figure A5.2 Comparison of models developed from three data points compared to four data points- for Reading, 2008



A thicker line is drawn for the 'All data' case where all four data points are used to fit the curve. For this initial exploration the Test score at age 0 is taken as 0, and it included as an additional point. The impact of alternative assumptions about the 'best' assumptions for the Test score at age 0 is explored in a subsequent section. The purpose of this section is to understand the effect of missing data by systematically leaving out known data.

The 'All data' curve is regarded as the best estimate of the trajectory for reading growth since it draws on all available data (that is all four points at 8.5, 10.5. 12.4 and 14.4). Ignoring data points that are not at the highest and lowest age has little impact of the trajectory fitted for the data. The 'All data' curve and the curves that use three points that include the lowest and highest age (8 10 14, 8 12 14) produce almost identical trajectories in the range from age 6 to 15.

The two models where either the highest or lowest data points are not included produce slightly different trajectories as marked on the chart. Logically the case that deletes the highest point (8 10 12) tends to a lower asymptote. The case that deletes the lowest point (10 12 14) tracks generally lower than the 'All data' curve below age 10. However in the range of ages from 8 to 14, the models with the widest time separation (even if missing one data point) produce very similar trajectories.

The chart also illustrates why the Gompertz model is attractive as a model for the general trajectory. The asymmetrically placed inflection point offers a possible explanatory mechanism for what may be happening (at a group level with quite a spread of performance) with reading development with age. Accepting the sigmoid shape as a reasonable general

model, and the asymmetric Gompertz curve as an appropriate choice, a model for the 'rate of learning' is provided. The curve of the annual rate of learning at each point on the trajectory for 'All data' is shown and scaled on the right axis. The rate is increasing as the inflection point (at about age 6) is approached from the left. On the curves fitted to the 2008 reading data the rate of learning is increasing rapidly for the ages from 3 to 6, peaks at about age 6 (at about 66 scale points per annum) and then reduces from ages 6 to 15.

There is however a potential impact on the fit of models from the choice of an appropriate value for the 'notional' test score at birth (age=0). The test scale zero is arbitrary and it is assumed that the NAPLAN developers gave no consideration to the choice of an 'absolute zero' for learning. If the Gompertz model applies, the degree of fit is markedly influenced by the choice of Test score value at age=0. This issue is considered in the following section as part of the development of a general model for reading learning trajectories.

Test scale transformation effects on a model for the trajectory of NAPLAN Reading 2008

In developing a likely trajectory for reading, the impact of alternative placements of the zero position for the test scale at age = 0 were investigated under the Gompertz model. Figure A5.3 shows the effects of various placements of the origin on the location of the fitted curve. Each case is a displacement from the original position and then the rescaled data are fitted to a Gompertz model. Once the model has been fitted, the data are rescaled to the original scale. As an example consider the case of OS-100. The data are rescaled by subtracting 100 from each of the original four Test scale means, fitting the new values to a Gompertz model, calculating the model value for each age from 0 to 18 and then adding 100 to these values. With this procedure the 0 position on the original scale was raised 100 units, the curve fitted and then the points were recalculated relative to the previous origin.

Figure A5.3 Effect of changing the location of 0 on the Test scale at age =0

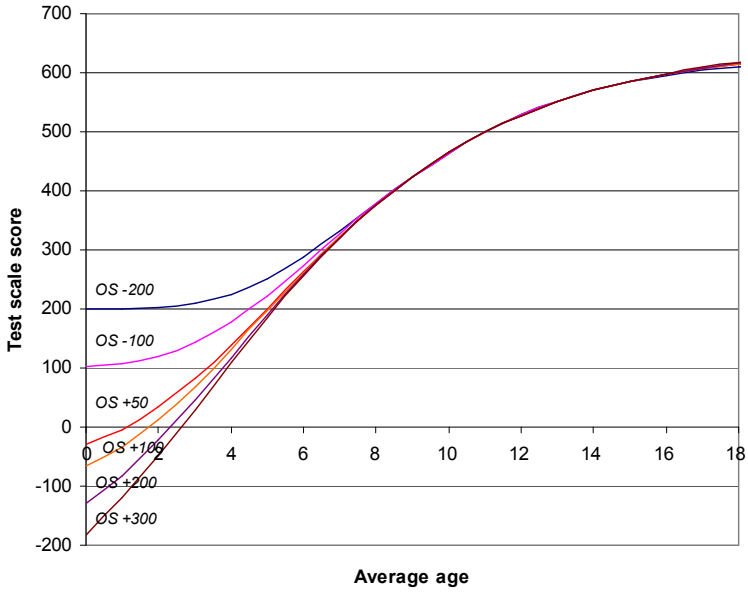


Figure A5.3 shows that the curves of all models are effectively coincident from age 7 to 16. The variation in the Test 0 position influences the shape of the trajectory from age 0 to 7. It does this by repositioning the inflection point. The higher up the scale the 'real 0' is assumed to be, the higher the inflection point. For OS-200 (Original scale minus 200) the inflection point is close to 8 years of age, for OS+300 the inflection point is around 4 years. The curve past the inflection point remains essentially the same, whatever the assumed Test zero. This exploration establishes that while the assumed zero position is important in determining at

which age the inflection point occurs, the trajectory after this age is essentially the same for all cases within a range of 250 scale units above or below the original zero position. This means that the curve segments for the ages for which data are obtained by tests are essentially the same. The choice of origin influences the steepness of the curve below age 8 and the model determined point at which the rate of learning peaks. This insight implies that the match of the extrapolation of learning status for younger ages in a Gompertz model is problematic but adds some general considerations for hypothesis testing.

A reconsidered model based on the original test scale

Given the insight above that the selection of the zero position for the test scale in the Gompertz model influences the location of the inflection point, the selection of the appropriate value for the test scale at age=0 should be based on the assumption (or known reality) of the likely age for the highest 'rate' of learning. A reasonable assumption for this period is in the first year of school, the ages of 5 to 6. As it turns out the NAPLAN scale, unchanged, fitted to a Gompertz model, has its highest rate of learning in the period from age 5 to 6. For this reason the further refinement of the model to match data points for Australian students is based on the original test scale. As indicated earlier the purpose for doing this is to provide an understanding of what cohort growth appears to look like in Australian school systems, to help select parameters for extrapolating data as outlined earlier is addressed in Chapter 6.

While the original scale is used, the value assigned to the test scale at age = 0 remains a matter of choice. The effect of the value of the test score at age = 0 can be considered through varying the value for the test score at this point, in a narrower range than explored above, to see if there is a optimum position for the data point at time 0, i.e. age = 0. Figure A5.4 shows the effect on the parameters of the fitted curves, on the standard error of the fit and the correlation of data points with the model, as the value on the test score axis at age 0 is systematically varied.

A matter to clarify is where the Gompertz model would place the value of the test score at age=0. In seeking a fit to all four data points (8,10,12,14), the modelling software iterates a solution that has parameters a, b and c as 642.1, 1.38, 0.251 in four iterations, with a standard error of 2.66. By adding a range of fifth points (age=0, and test scores varying from 0 to 60) model fit can be optimised by author-managed alternative models to minimise the standard error and maximise the coefficient of correlation (R^2). While the four data points (ages 8,10,12,14) are fitted to the model in four iterations, it is possible to find a value for a fifth point, the test score at age zero, that reduces the required iterations to 3 and minimises the standard error of the fit of the data to the model, and maximises the correlation of coefficient of the fit. This point is found at Test scale =12. Figure A5.4 panels show the variations in the parameters as the test scale value at age zero is varied. With a value of 60, the asymptote (a) increases to over 700, relative to a value of 623 at Test scale =0.

In Figure A5.4 the model with Test scale =12 has an asymptote at 642 the same as the four points only solution. The effect of the placement of the fifth point at Test scale = 12 is to reduce the iterations to solution and to reduce the standard error to below 2. While the 5th point is possibly unnecessary, the optimised Y-axis intercept is close to the value of 13 derived in Figure A5.2. As noted in that case, a model based on all states and the ACT (but not including the NT) has an asymptote of 649 (648.97) compared with 642 above and a model-selected intercept of 13. The model based on the national averages only, has parameters a, b, c and a Y- intercept close to those resolved in Figure A5.2.

Based on the Test scale value of 12 at age 0, and the four test scale mean points, a general model of the national age cohort trajectory for reading can be plotted as illustrated in Figure A5.5.

Figure A5.4 Panels of Parameters and Fit indicators for Options for Test score at Age=0

Gompertz Relation expression: y=a*exp(-exp(b-c*Age))

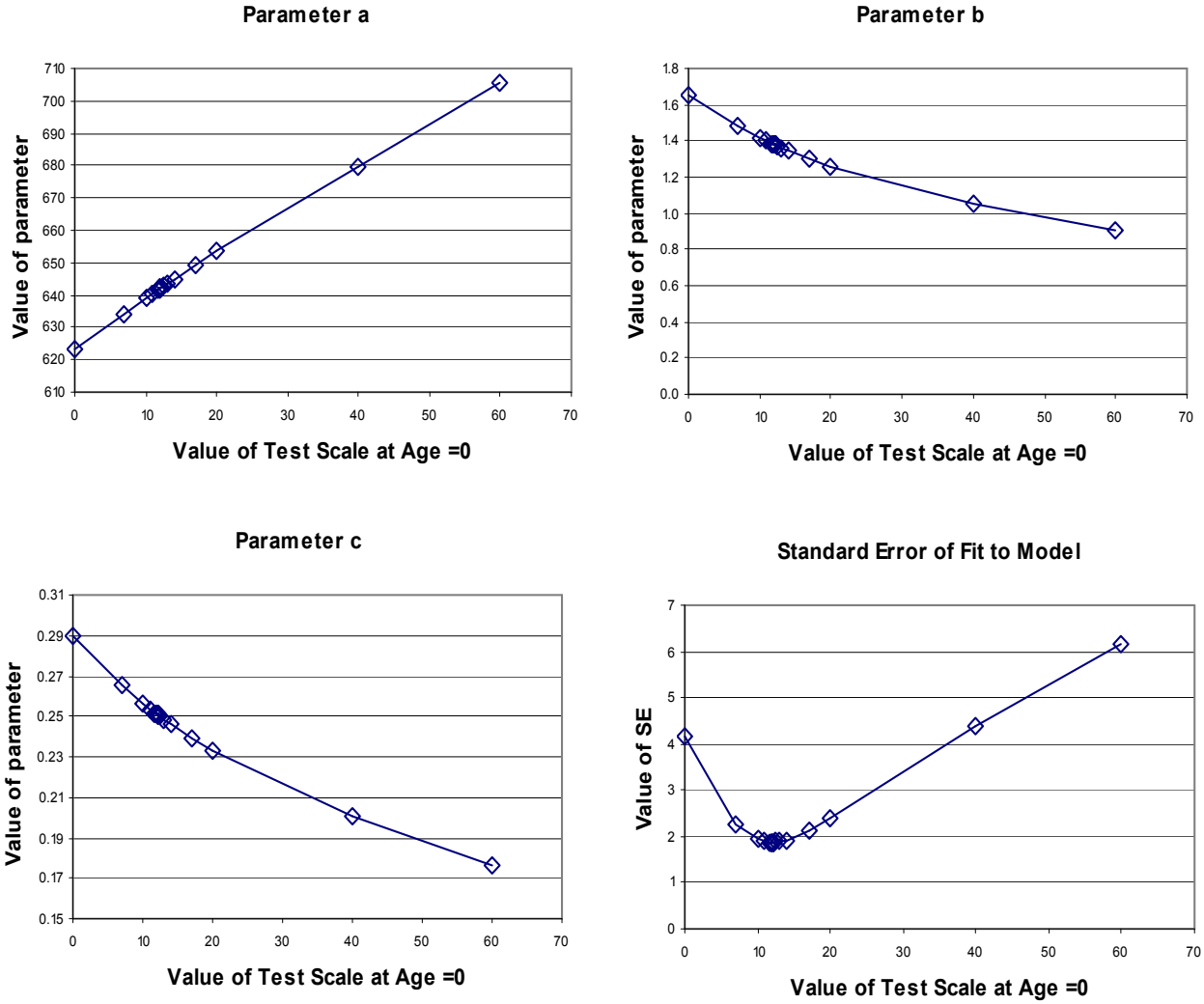
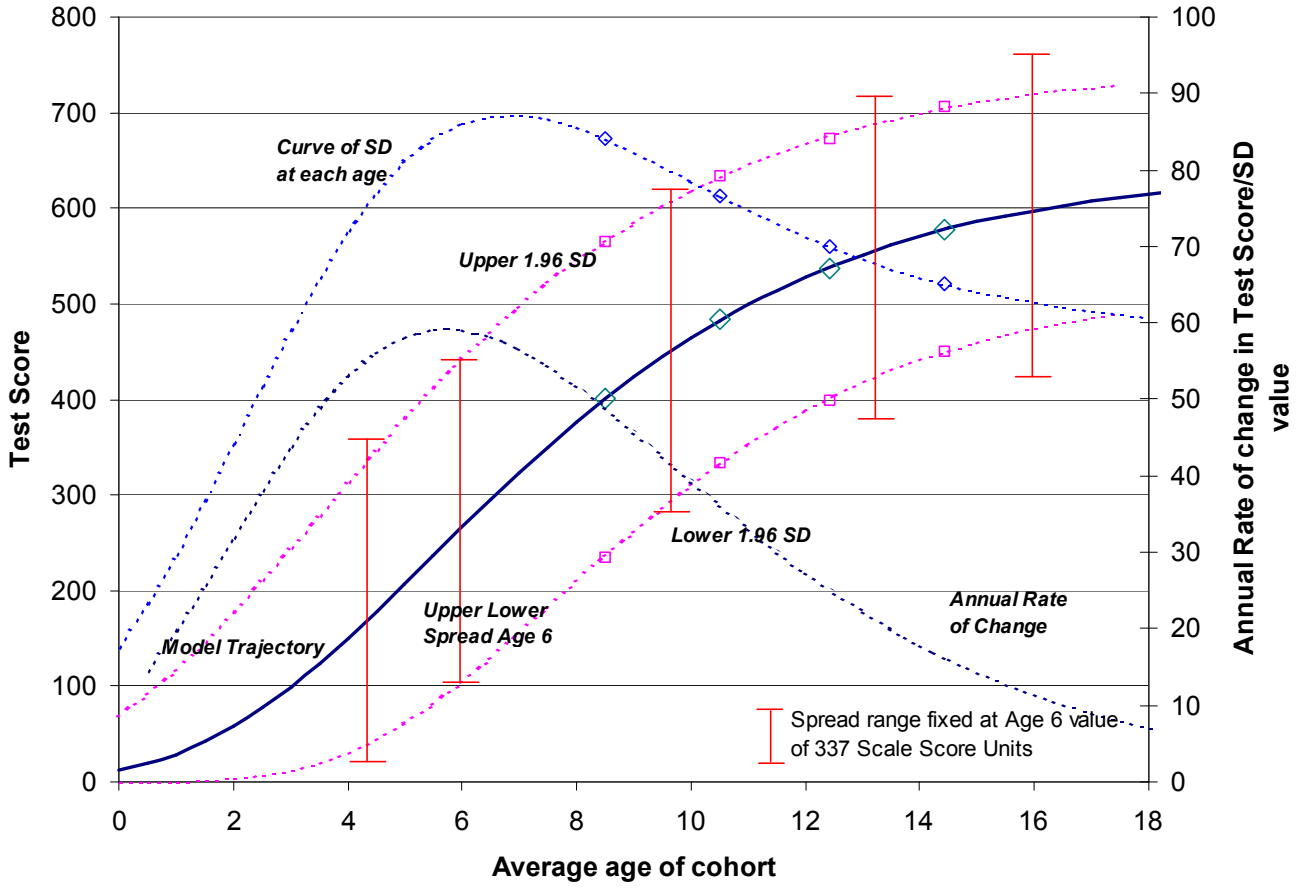


Figure A5.5 Model of NAPLAN Reading 2008 with indication of spread of data



The model in Figure A5.5 illustrates the annual rate of change in reading development at each point on the age scale by comparing reading status at successive age points. Estimates of the general patterns of the SDs can be made. The points at 1.96 SDs above and below the four national means encompass 95%of the student scores at each age. Curves can be fitted to the four actual points that delineate these upper and lower boundaries of the 95% of cases (assuming a normal distribution at each age point). The upper curve has an asymptote at 751 and notional test score intercept at age = 0 of 68 test scale units. The lower curve has an asymptote at 521 and a notional test score intercept at age = 0 of 0.08 test scale units. By subtracting the lower boundary values from the upper boundary values an estimate of the SD at each age point can be made by dividing the resulting value by 3.92 (2 x 1.96). On this basis, estimates of the SD can be made for any age points, enabling as a consequence the estimation of the effect sizes for annual growth at those age points (assuming n as the values used in generating the data in the first place)[35].

The resulting SD estimates are plotted with their scale on the right hand axis. The estimated SDs start small, grow to about 87 test scale units at about age 7 and reduce from that point on. Based on the observation of Schulz & Nicewander (1997) that growth spurts (eg puberty for

[35] In a model where all students are spread on the age axis at their actual age at testing the general distribution of data points remains essentially the same. The value of n for any point is the cohort n divided by the number of age categories for the cohort. Since values of n above 50 make little difference to the effect size calculation the actual n is almost immaterial.

human height) lead to greater variance at that spurt point, it is confirmed in this model that the SDs are greater around points of rapid growth, that is near the inflection point. The peak SD lags the peak rate of learning development by about a year. The peak rate of learning is around 6 years, the peak SD around 7. In this model logical and mathematical reasons are provided for the 'scale shrinkage' (Yen, 1986; Camilli, Yamamoto & Wang, 1993), the shrinkage of SD within a year level (very small) and more observable, the reduction of SD at higher Year levels.

The estimated annual rate of learning at each age is also plotted on the right hand scale. This curve illustrates an implication of the model. Students near the trajectory of the mean are likely to be learning at their maximum rate about age 6.

On the basis of estimating the SDs from the modelled upper and lower bounds, Figure A5.5 reflects a model of what the data might look like if students were assessed by an appropriate process at all ages from 0 to 15 on the one day and those data plotted by the average age for the students, aggregated to cohorts of average age, say in 0.1 decimal years of age at testing. An alternative view is that if all student scale scores were plotted individually by actual student age at assessment as individual points, 95% of these points would lie within the upper and lower bounds assuming a normal distribution by age on the test scale.

The actual data points reflecting the upper and lower bounds based on actual data points are identified on the curves, along with the actual scores on the model trajectory and the actual SDs.

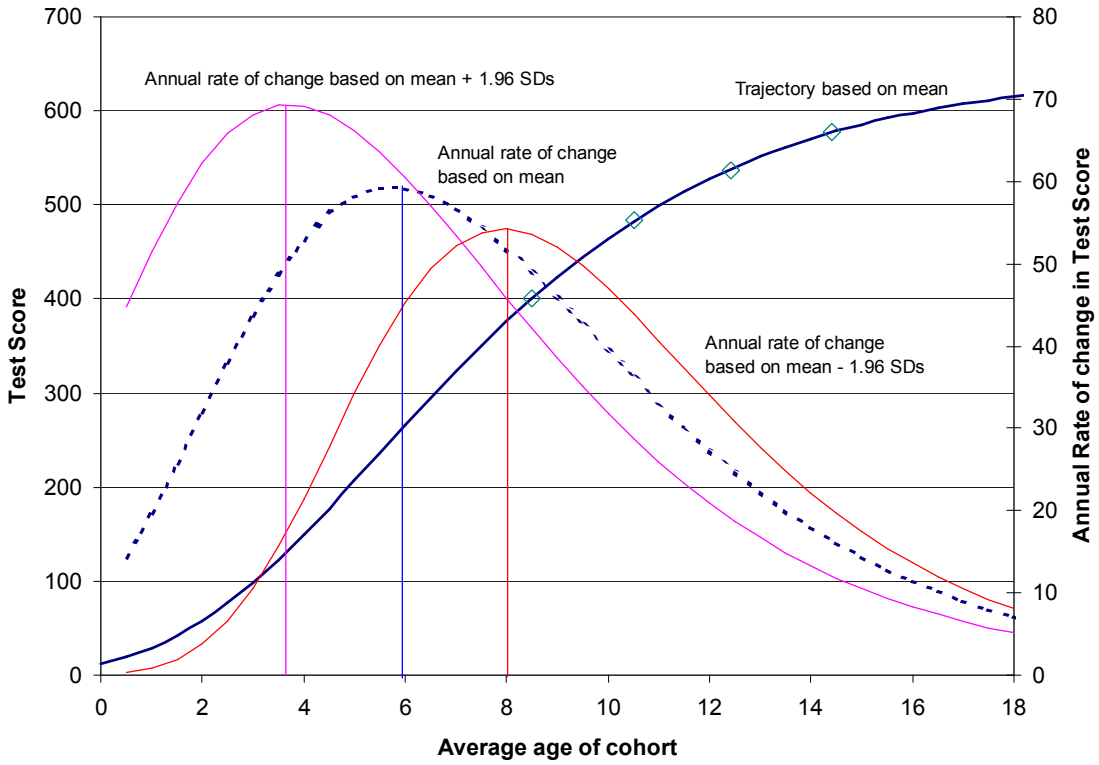
The bars on the chart are all of constant length, based on the estimated SD at age 6, and indicate visually the reducing spread of the scores at higher ages. Based on the explorations of test scale transformations above, speculating about what happens below the inflection point will be inaccurate in the estimate of the actual trajectory from age 0 to 6. A diminishing SD is however plausible because the rate of growth is smaller and the actual quantum of learning that is possible is less. A better basis for estimating the range of 'pre reading skills' would provide a better model. The author has allowed the trajectory to start at age=0 for completeness. It can be assumed that at some future point a better understanding of learning of the appropriate skills in younger children could be incorporated into such a speculative model.

Further speculative implications of the NAPLAN Reading model

The model can be explored further by plotting the annual rates of learning for the mean in comparison to the rates implied in the upper and lower boundary curves.

Figure A5.6 shows these three speculative annual rates of learning. The small number of students near the upper boundary are learning at a high rate and peak at just below 4. The group near the average have a gentler increase in rate and peak at 6. The students near the lower bound start slowly and reach a peak at age 8. Surprisingly the model suggests their rate of learning continues at a higher rate than the mean group or the upper tail. Were this simple model anywhere near a model of reality, some hypotheses about individual learning trajectories could be tested using individual longitudinal data. The issues relating to individual trajectories are addressed briefly later in the Chapter 5 and Appendix 10.

Figure A5.6 Model of NAPLAN Reading 2008 with annual rates of change in learning at mean, 1.96 SDs above and 1.96 SDs below the mean



While the model illustrates in an approximate way what data might look like if say all students were assessed at the one point in time and their data plotted, the model can also estimate what the mean score for a cohort at a particular cohort average age might look like. Using the model in this way enables the effect size for 'normal' year-to-year growth to be estimated.

Impact of data spread on State comparisons

A specific insight from the curve-fitting with average age at testing is the impact of this placement of the test mean on comparisons between states and territories (or school systems generally), where the average ages for systems vary markedly from each other or from the national average age. A similar situation applies when the data are analysed by elapsed years of schooling rather than age. Both differ from the impression gained when 'age' is centred on Year level. Table A5.1 lists the grand means of the displacements from the national test score mean at each year level for each system, averaged over the four Year levels, compared with the displacement from the fitted curve (Gompertz).

Table A5.1 NAPLAN Reading 2008 – Comparison by State and Territories by displacement from National mean

	Mean Age at Year 9	Mean displacement from National mean (averaged over all four Year levels)	Mean displacement from curve (averaged over all four Year levels)	Difference (Curve Displacement minus –Nat. Mean Displacement)
Qld	14.08	-16.48	-0.57	**15.90**
WA	14.00	-10.58	-0.27	**10.30**
NT	14.42	-73.78	-70.85	**2.93**
SA	14.50	-3.15	-2.29	**0.86**
NSW	14.58	8.30	8.33	**0.03**
ACT	14.67	21.25	18.77	**-2.48**
VIC	14.75	11.20	6.24	**-4.96**
Tas	14.83	-2.20	-11.23	**-9.03**

The table is ordered by the impact of the difference between the two methods of establishing displacement. Queensland is displaced from the national means at each Year level by an average of -16.46 test scale units. Apart from the Northern Territory this is the largest negative (below mean) displacement. However when the mean test scores for Queensland are compared with the curve of mean test scores by average age at testing as the X axis, as shown in the right panel of Figure 5.1, the displacement from the curve is less than one test scale unit. Effectively when an adjustment is made for the lower average age in Queensland at each Year level, Queensland sits almost where it would be expected to be, accepting the plotted curve as a reasonable model for scores by average age. Similarly Western Australia is almost where it would be expected (-0.27 units from the curve). NSW and SA maintain the same relationship with the curve model as they do with the national mean (due to the closeness of their average ages to the national average age). The ACT's displacement is slightly reduced as the average age is slightly higher than the national average. Victoria, which on the national mean score comparison is 11.2 units above the national mean, is only 6.24 units above where the model by age would place it. Tasmania, which appears to be close to the national mean, is actually 11.23 units below the curve when age is considered. The Northern Territory being at the national average age is unchanged in the comparison. Whether or not the model suggested is the most appropriate, the principle is established that an age adjustment when making national comparisons produces quite different conclusions.

Appendix 6 North West Evaluation Association -Data confirming learning trajectories

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

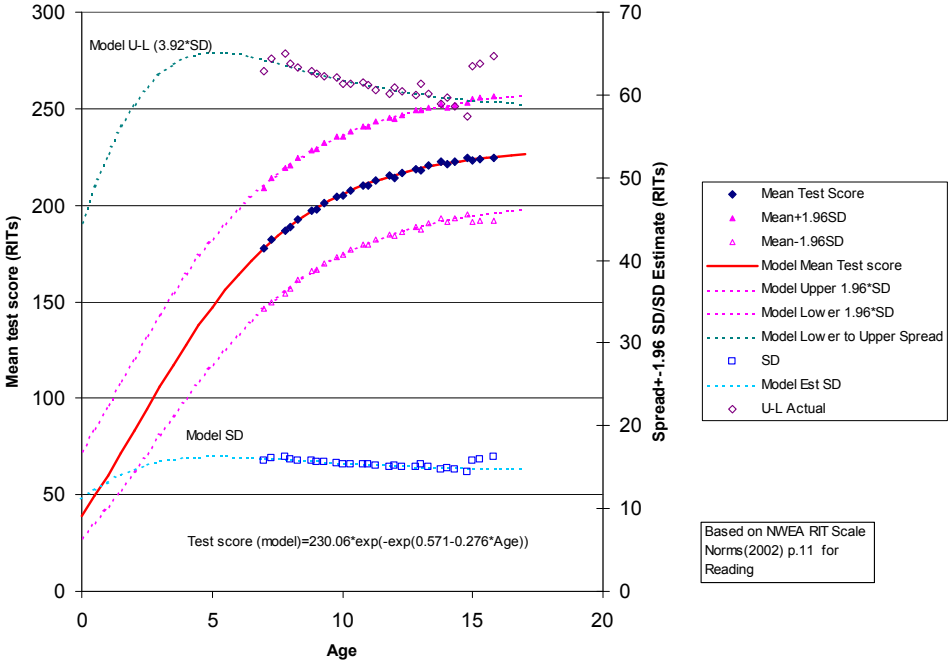
Most of the tests are adaptive and are vertically scaled, drawing on an item bank of 15,000 items. Tests are completed at a computer screen and the process adjusts the difficulty of the items to the current ability of the student. As a result the tests are grade independent. The scale is equal-interval,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

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

Growth Trajectory based on NWEA norms

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

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



To spread .3
(Winter), ´ .0

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

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

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

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

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

Appendix 7 Mathematics Assessment for Learning and Teaching (MaLT) in England

Williams, Wo, & Lewis (2007) and Ryan & Williams (2007) report data from a national sample designed to provide age related performance references for the MaLT test. Year level cohorts of between 1000 and 1400 students were recruited from 111 schools.

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

A scatter plot of all students across all age categories, in months is presented as a 'Quintic' model in Williams et al. (2007, Figure 1, p. 134) with test scores reported as logits. From this figure it was possible to estimate some broad values to develop a model as illustrated in Figure A7.1. The data in this model are estimated from the published scatter plots rather the r taken from tables.

The data are presented in Williams et al. (2007) as 5 trajectories representing the paths of the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentile students. To develop a model similar in form to those in previous appendices, readings of data points from the 50^{th}, 90^{th} and 10^{th} percentile trajectories were made. The assumption was made that the median was very close to the mean and the median then used as a proxy for the mean. The readings for the medians at a set of ages (60, 72, 84, 96, 108, 132, 168 months) were taken. Readings were taken for the same ages on the 90^{th} and 10^{th} percentile trajectories.

A model was estimated through curve fitting using the Gompertz equation for the three trajectories. To achieve this the logit scores were transformed to a scale with 0 logits = 200 and one logit transformed as 20 scale units. Once the model was fitted the results were reconverted to the original logit scale. As in previous appendices, the impact of the choice of the zero position of the scale score influences the shape of the 'modelled' trajectory below age 5 but with little impact above. Transforming the original scale with zero at 200 provided an inflection point at about 5 years of age. Transforming the original scale with zero at 400 produced an inflection at about 3 years of age.

The trajectories of the upper and lower boundaries for 95% of the data were estimated by first developing the model for the 90^{th} and 10^{th} percentiles, establishing the spread between the two lines at age points, re-scaling the area under the normal curve from 80% to estimate a SD (by dividing by 2.56 (2*1.28)). The upper and lower boundaries are then estimated from 1.96 * SD.

Figure A7.1 displays the resulting model for the mean, the actual data points (as estimated) and the upper and lower boundaries for 95% of the data. Also plotted are the annual rate of change, based on the model and the estimate of the SD. Consistent with the NAPLAN model the model SD reduces slightly as the age increases and peaks about a year of age past the inflection point. As previously the model estimates can be used to estimate the effect sizes for year-to-year growth. These are shown in Figure A7.2. For reference the US effect sizes from Hill et al. (2007) for 6 Mathematics tests are included (taken from Chapter 5, Table 5.1).

Figure A7.1 Model of Mathematics Development - Mathematics Assessment for Learning and Teaching, (MaLT)

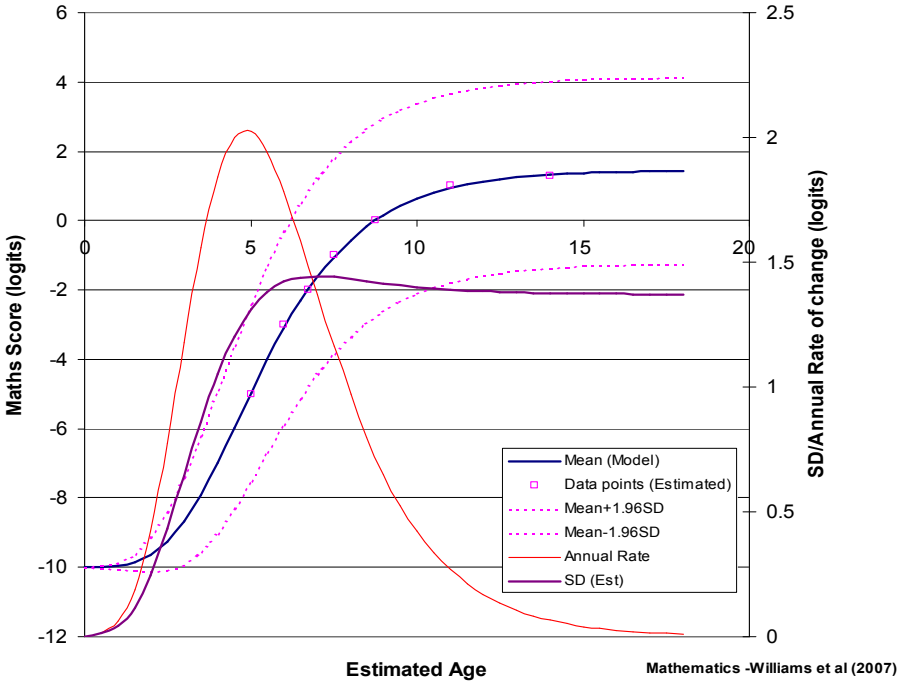
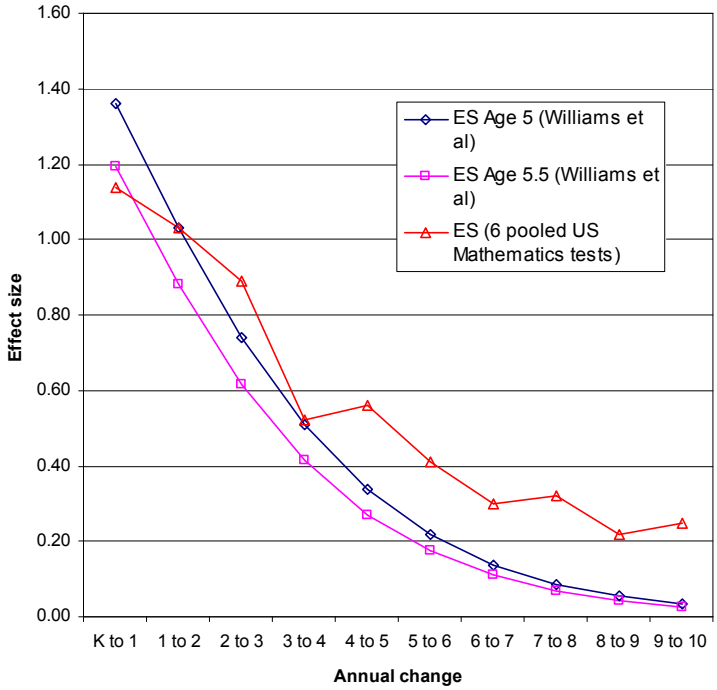


Figure A7.2 Effect sizes for Mathematics Assessment for Learning and Teaching compared with pooled US tests.



The general pattern of reducing learning growth in mathematics is exhibited, through reducing effect sizes, in both the Williams et al. and US data. Somewhat surprisingly, for the UK model, the growth in logits per annum from Figure A7.1 and the effect size in SDs from Figure A7.2 are both close to zero by the transition from Year 8 to 9, much lower than the US comparison. (The Year 10 to 11 effect in both figures is an extrapolation- Williams et al. only tested to Year 9). It was this plateau effect that was the focus of the Williams et al. (2007) article since it implies almost no mathematics development from Years 7 to 9.

The validity of the vertical scale is considered in the article but even with the qualifications to the phenomenon that might be related to the inadequacy of the scaling, Williams et al. conclude that even though the

> …the plateau … must be interpreted with the limitations of the vertical equation methodology in mind …. closer examination of the three year model with superimposed one-year models confirms that the plateau is not an artifice of the full 10 year vertical equating model. It seems realistic to conclude that progress is indeed very slow (about 0.2 logits per year) over this period. …… One speculates that the repeated exposure to the same curriculum in secondary school has a negative effect on these common learning outcomes. (Williams et al., 2007, p. 139)

The Williams et al.(2007) data are also helpful in illustrating the age effect within a Year level cohort. This phenomenon appears to apply quite generally and is discussed in Chapter 5.

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.

Tymms & Wylde (2003), Tymms, Merrell & Jones (2004), and Merrell & Tymms (2007) published data on cross-cultural difficulty patterns for sets of Mathematics, Reading and Vocabulary test items in teacher administered, computer-adaptive tests using the Performance Indicators in Primary Schools (PIPS) On-Entry Baseline Assessment. The PIPS On-Entry Baseline Assessment, developed and managed by the CEM centre, combines objective assessment and teacher ratings to provide information about each child as they enter their first year in full time education. At the core of the PIPS On-Entry Baseline is an assessment of early reading, early mathematics, phonological awareness and short-term memory. The assessment is completed by an adult (usually but not necessarily the teacher) working with each child on a one-to-one basis at a computer screen and takes about 20 minutes.

As a result of the very comprehensive record storage process, including the individual responses at each assessment session, a database has been developed covering a large number of cases. The database has records across countries including England, Scotland, the Netherlands, Germany, Hong Kong, Australia and New Zealand.

Merrell & Tymms (2007) reported student responses from England, Scotland, Western Australia and New Zealand. Although English students could also be tested in Bengali, Cantonese or Urdu only the English language tested data subset was used (Tymms, personal correspondence, 2008). Other publications (Tymms & Wylde, 2003; Tymms et al., 2004) include students from the Netherlands (in Dutch), Western Australian Indigenous communities and English hearing –impaired students. Tymms has also supplied the detailed item difficulties for some of the samples involved, to the author (Tymms, personal communication, 2006). All these sources have been drawn upon to develop a conceptual argument concerning the possibility of invariance of item parameters across cultures in the learning of numbers in English.

If certain numerals appear to be recognised before others, this phenomenon allows the observer to monitor learning progress as more difficult numerals are recognised. More importantly, if the 'distance' from demonstrating the ability to recognise a particular numeral is consistently the same learning distance (in terms of differences of item difficulty) from another specified numeral, there are strong hints that there is a scale for the learning of numerals. The conditions of order and consistent intervals, the prerequisites for measurement, are met.

The Tymms et al. data provide convincing examples of consistency of item order across cultures, strongly suggesting in numeral development at least, there is a natural approximate order in which students master the naming and recall of numerals 0 to 9, that is in their development of a language of number words. Extending this further to naming 2 and 3 digit numerals and computations, the consistent item difficulties (and inter-item interval distances) obtained, are shown to be independent of the English-speaking culture from which a student is derived.

More broadly Merrell & Tymms (2007) reported across a wider set of items covering mathematics, reading and vocabulary. The strongest correlations between item difficulties across countries were in mathematics. Correlations of the difficulties of the items in the four countries (England, Scotland, New Zealand and (Western) Australia) were all 0.99. "This is so high that no further preliminary action was needed before making comparisons. The difficulties of the reading items were also strongly related but not quite so strongly." (Merrell & Tymms, 2007, p. 127.). Table 8.1 reproduces Tables 6 to 8 from Merrell & Tymms (2007)

showing the inter-item difficulty correlations across countries. The reading correlations (Table 6) ranged from 0.99 to 0.94, vocabulary items (Table 7) ranged from 0.99 to 0.95 and phonological awareness items from 0.91 to 0.98.

Table A8.1 Correlations between item difficulties, by country, reprinted from Merrell & Tymms (2007)

table 6 correlation between difficulties of 56 reading items[a]

	WA	NZ	England
NZ	0.99		
England	0.97	0.96	
Scotland	0.96	0.94	0.99

table 7 correlation between difficulties of 17 vocabulary items[a]

	WA	NZ	England
NZ	0.96		
England	0.97	0.96	
Scotland	0.95	0.96	0.99

table 8 correlation between difficulties of 17 phonological awareness items[a]

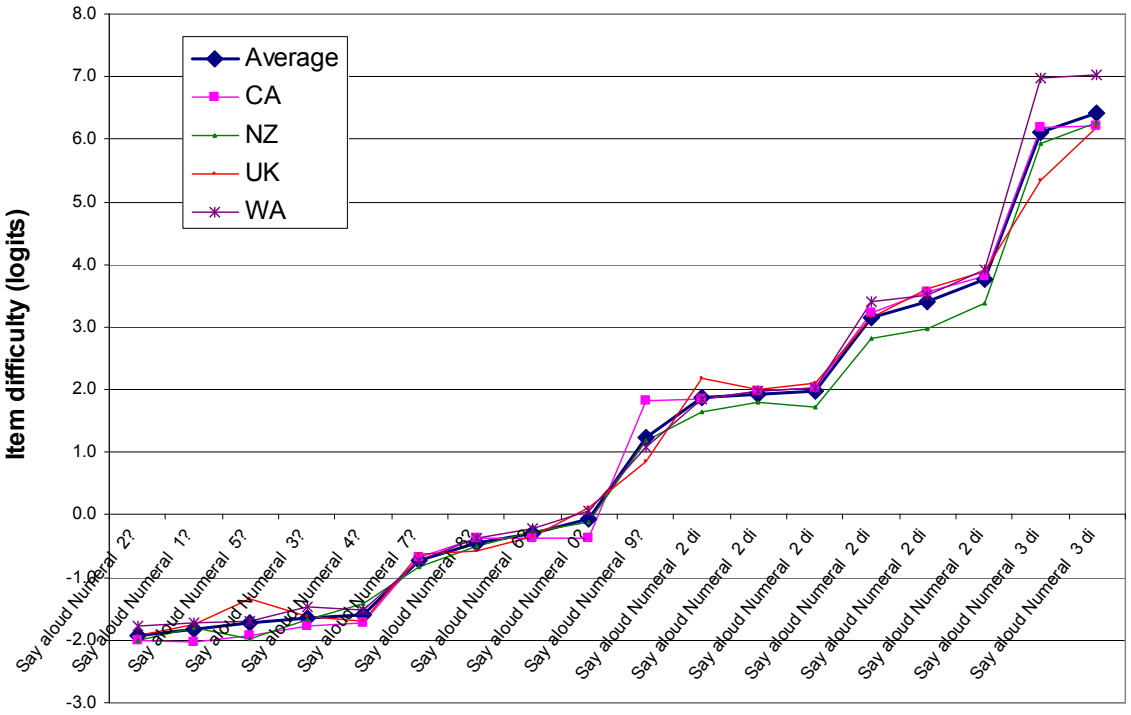
	WA	NZ	England
NZ	0.98		
England	0.96	0.98	
Scotland	0.91	0.91	0.95

[a] A few infrequently presented items with large errors were omitted.

Figure A8.1, based on the data supplied by Tymms (personal communication, 2006), illustrates that there were minimal variations in individual item difficulties across cultures in saying aloud numerals. English speaking Cantonese (CA) show a slight advantage (easier to say the number name), consistent with Miller, Smith, Zhu & Zhang (1995), where the simplicity of Chinese word names provide an advantage to Chinese speakers. This advantage would appear to flow over to English language names.

The line of mean item difficulties (designated as 'average' in the figure) shows items in increasing difficulty order; the individual sample lines illustrate the small variation in difficulty across cultures. New Zealand for example appears to show a slight advantage (i.e. easier) for naming two digit numerals. Cantonese English speakers learning English numerals found 9 harder to master than did other language/cultural groups. Three digit numbers were harder to master in Western Australia than elsewhere. The pattern of similarity across cultures is, however, remarkably consistent.

Figure A8.1 Item Difficulties over Four Culturally Different Samples (Tymms, personal communication, 2006)

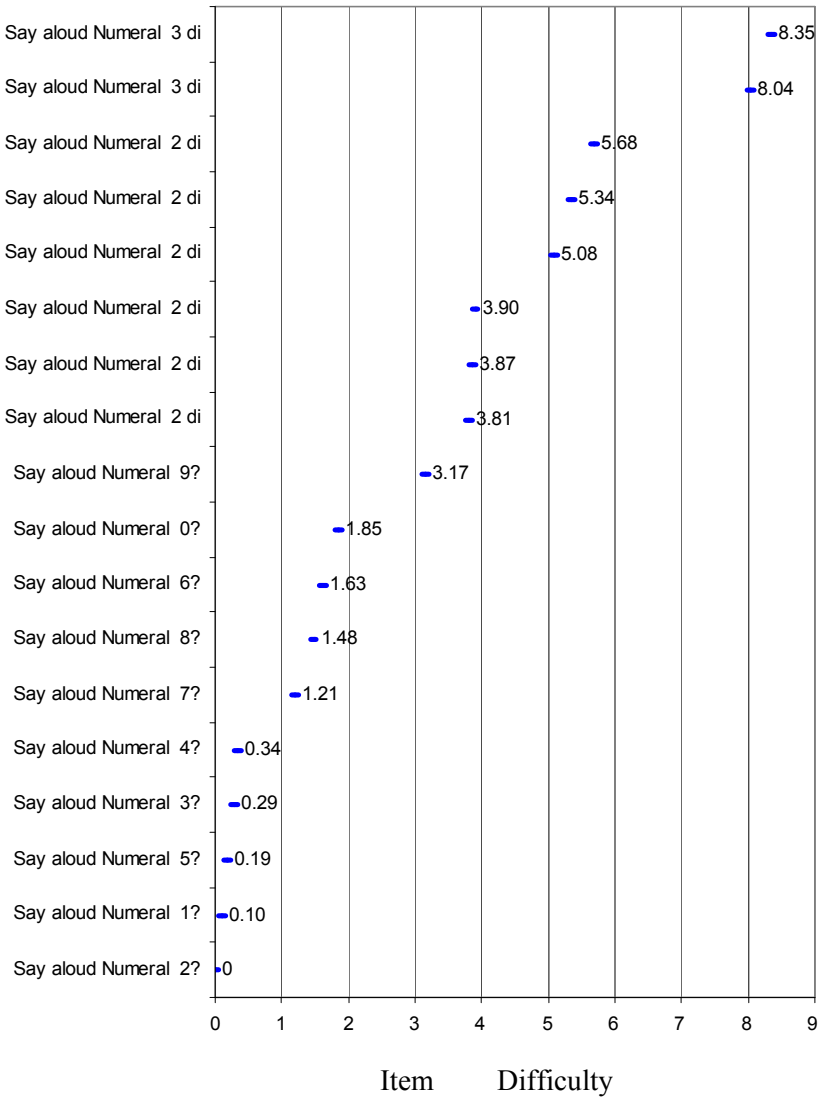


Tymms et al. (2004) and Merrell & Tymms (2007) applied differential item function (DIF) analyses to show that there was a no difference (i.e., lack of bias) in numerical recognition items, for any of the cultural groups. Some DIF was found for a small number of vocabulary items and explained as culturally related (`wasp' and `pigeon' were more difficult in Australia). These estimates of student performance also show a strong pattern by age, similar to the NAPLAN data. .

In more detail, Figure A8.2 presents a map of the item difficulties, anchored to the difficulty of learning to recognise the numeral 2 (say aloud the number word, the marginally easiest numeral to identify over all cases[36]).

[36] Later in this Appendix, a similar analysis, but with more cases, reverses the position of 1 and 2.

Figure A8.2 Map of Item Difficulties anchored to Recognise/Say Aloud 2 =0



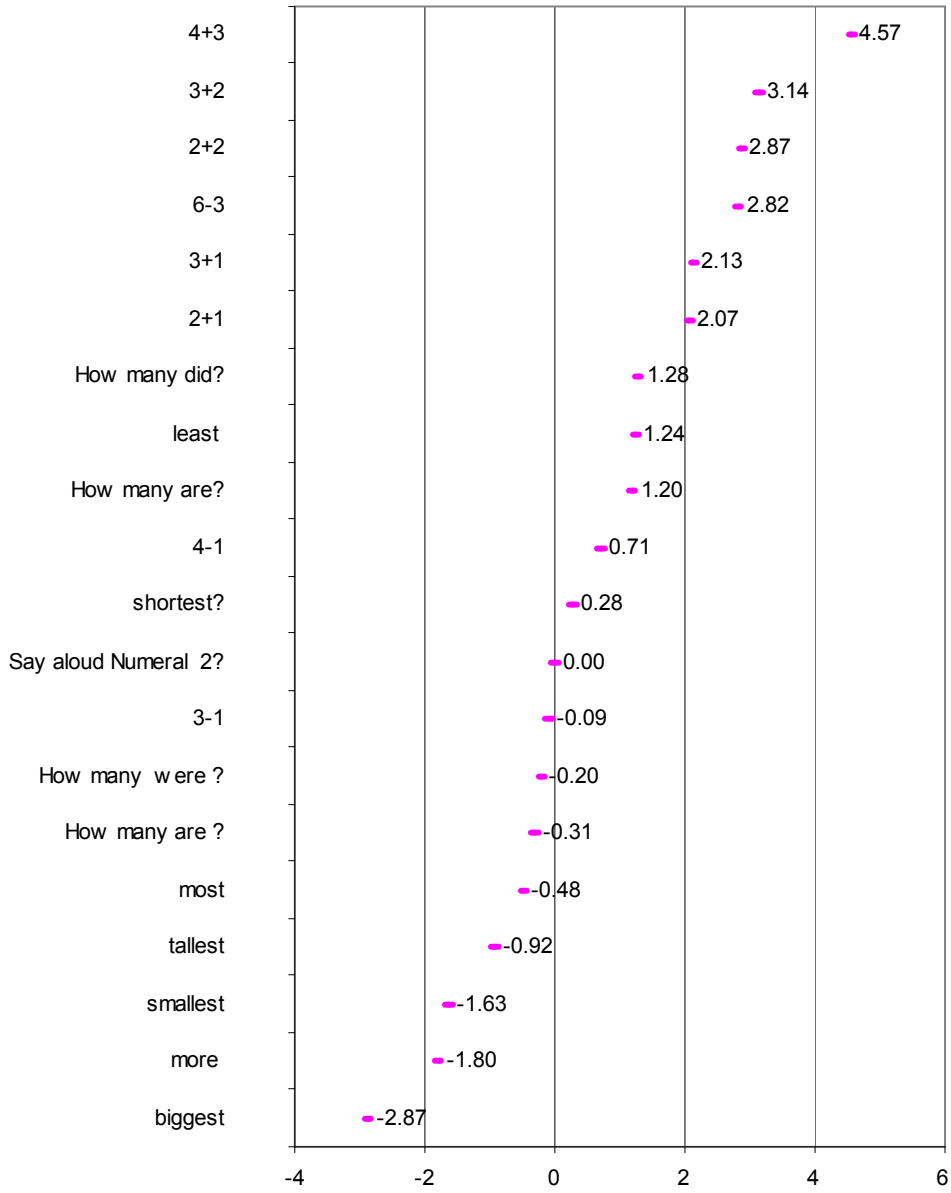
The CEM data set is expanding annually and internationally and thus more data might revise the order. Based on the data to date there is an order in numeral recognition which is plausible (that is logically consistent with experience) but at a level of refinement apparently not appreciated by most observers. It appears 2 is recognised marginally ahead of 1, that is it is slightly easier to recognise. The order of numeral recognition appears to be 2,1,5,3,4,7,8,6,0,9. The increase in difficulty from recognising 4 to recognising 7 is almost one logit, the distance from 2 to 9 being over 3 logits. The challenge to young children in building numeral recognition skills (as a small example of the complexity of all the early mathematics and language skills) is great. The change in difficulty levels of average Year 3 students progressing to Year 5 in mathematics or literacy is approximately 1 logit, although the logits may not be directly comparable. The high rate of change of learning development implied in early number recognition is however consistent with the diminishing learning rate with time described earlier in Chapter 5.

From the map in Figure A8.2 saying aloud 3 digit numerals, by implication, the recognition of place value using appropriate descriptions is more difficult than recognising numeral 2 by more than 8 logits, a large increase in difficulty. Average performance of students measured in logits (though not exactly comparable) in progressing from Year 2 to Year 12 is estimated to be about 6 logits.

Figure A8.3 maps the difficulties of small calculations and the recognition of relativities (larger, smaller, most), once again anchored to a difficulty of recognising 2 given a value of 0. Identifying which is the biggest of three items (cats) is almost 3 logits easier than recognising 2. Counting 4 items is about 0.3 logits (i.e., just measurably) easier than recognising the numeral 2.

Counter-intuitively, but not overly surprising, subtracting '1' seems to be easier than adding 1. The calculation '3-1' in the form of 'Here are three beach balls. If you took one away how many would be left?' has a difficulty of –0.9 (i.e., it is easier than recognising the numeral 2) while the sum '3+1' in the form, 'Here are three bikes. If you put one more bike in the picture how many would there be?' has a relative difficulty of 2.13, over 2 logits more difficult than the subtraction equivalent.

Figure A8.3 Map of Difficulties for Relative and Computational Items (anchored to 'Say aloud 2' =0 logits)



The subtraction '4-1' is easier than the sum '3+1' and the subtraction '6-3' easier than '4+3' by 1.6 logits. Meanwhile relative terms 'shortest' and 'least' appear to be much harder than their opposites, 'tallest' and 'most'.

Additional data (Merrell and Tymms, 2007), not detailed here, illustrate similar regularities in the areas of reading, phonological awareness and vocabulary. The full suite of school entry assessment items covers writing, vocabulary, ideas about reading, repeating words (assessment of phonological awareness), rhyming words, letter identification, word recognition and reading, ideas about mathematics, counting and numeracy, sums (addition and subtraction problems presented without symbols), shape identification, digit (numeral) identification (single, two and three digit numerals), and mathematics problems (including calculations with symbols).

The purpose of this analysis is to illustrate that developmental maps, based on empirical student-derived data could provide teachers with some of the tools to note and understand each student's progress. This is a key element of the general thesis, that teachers with the right conceptual tools (likely-order maps say) could observe, understand and record student progress. This proposition is not particularly earth shattering. As described in the main text, Masters and Forster (1996) and others have been proposing similar approaches for almost two decades. The advantage brought by the CEM data is the confidence it should give to teachers in classrooms and others, that the patterns of order of skill development are genuine and consistent across cultures and thus reflect some significant underlying characteristic of learning. Particular skills and developing abilities appear, unsurprisingly, to be dependent on earlier skills and abilities and progress along the developmental pathway is not necessarily smooth or easily achieved, based on the estimated 'difficulty measures' of learning later skills/tasks relative to earlier ones.

Item difficulty can be varied somewhat by the design of the item. A trivial and obvious example is the contrast of the difficulties of requiring the student to read the item (when the child cannot yet read) versus a teacher mediated computer presentation of the same item as CEM provides. In one form the computational or recognition skill alone is observed, in the other the computational skill and the ability to read are combined producing an item of much higher difficulty. Clearly establishing the relative difficulties of particular skills will require dissection of the contributors to item difficulty but the CEM data sets have already shown that the difficulty patterns are likely to be consistent over cultures and are thus most likely related to inherent properties of the particular cognitive skills relative to other skills.

Tymms et al.(2004) make a strong case for the usefulness of their assessment approach in cross-cultural studies to better understand cognitive development.

> The analyses presented … have explored the possibility that a baseline assessment (the PIPS On-entry Baseline) could be used to make comparisons of pupils starting school in different countries and cultures. The evidence suggests that this is indeed possible. The assessment behaved well across the groups that were studied and the general developmental patterns also appeared to be very similar across the groups. Clearly, some of the analyses indicate that more work is needed on the assessment items but that is to be expected in a pilot. The way is now open for a serious international study of the cognitive developmental levels of children starting school. (Tymms et al., 2004, p. 688)

A significant benefit of the CEM approach, through promotion of potential item maps derived from their data, would be to help teachers understand cognitive development. The cross cultural consistency should help teachers believe such skill development orders might be genuine and based on a roughly predictable model of learning development. The maps would become reference frames for understanding and documenting where each student is in real time.

CEM Number word development updated

More recent data (Jones, 2008, personal communication) is recorded in Figure A8.4. This new analysis determined the difficulty of every number presented from 0 to 999. Figure A8.4

is based on unpublished data from CEM. The difficulty scale is based on at least three separate analyses. To very approximately equate the scale for Figure A8.4 to be similar to Figure A8.2, the scale for Figure 8.4 was rescaled to make the difficulty difference from 1 to 9 to be 3.1 logits, the same as in Figure A8.2. The numerals follow approximately the same order but Figure A8.4 has identified each number sequence from 1 to 999, providing much more detail about the likely order of numeral learning, although for numerals higher up the order the differences in relative difficulty are too small to be meaningful. To illustrate the general trends, most 2 digit numbers in the thirties are about 1 unit more difficult than repeat digits (22, 44, 88), which are learnt it seems approximately in the order of the initial difficulty of their single digit components (66 and 99 exceptions). Saying aloud 100 is 2 units less difficult than 200, which is equivalent in difficulty to 101 the next hardest 3 digit number. The round hundreds are easier than most other three-digit combinations. The last numerals identified are 556, 716, 701, 770 and 917. The differences are small for a large portion of the difficulty sequence and the error of measurement is of the order of 0.75 to 1.0 units (a consequence of small numbers of cases combined with large individual differences in the difficulties of digits at the 'hard' end of the scale).

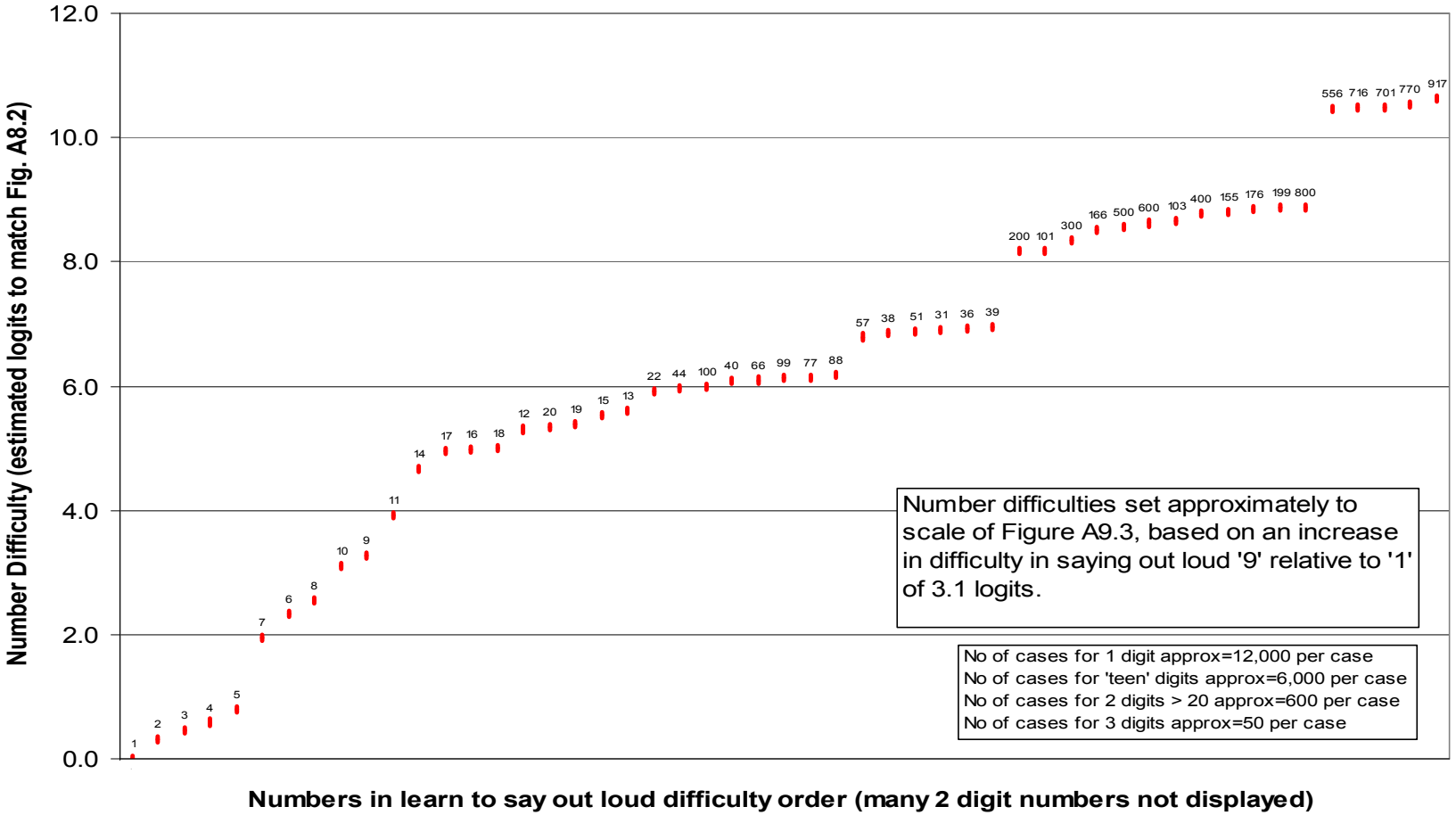
The link between developing a vocabulary of number words and the development of counting (which comes first?) has been investigated recently by Condry & Spelke, 2008. In a series of experiments with young children learning number words and counting they report their work

> provides evidence that natural number concepts emerge in children along with or after, rather than prior to, the acquisition of language. These concepts likely emerge, in part, as a consequence of children's efforts to make sense of number words and to learn to use the counting routine to represent number: achievements that the children in the present experiments had not yet attained. (Condry & Spelke, 2008, p. 35)

Surprisingly this is quite recent research. This understanding, combined with a number word-learning map of the sort derivable from CEM research and confirmable from other sources, would provide a teacher with an observation framework for number word development. A logit related scale would provide a basis for easy recording of learning status (and linking to actual testing at the next CEM assessment for clients of that system).

The CEM data are presented to illustrate why knowing the relative difficulties of learning to say aloud a number (recognise the digits and verbalise the name) might help a teacher. A student who can say aloud a single digit but not two digit number, can be recorded as having developed to position x at time t_1. Specific 2 digit numbers indicate position y at time t_2 and 3 digit numbers indicate position z at time t_3. The specific numbers verbalised are indicators of generally where a student is in their recognition of numbers. Assuming a more refined analysis than Figure A8.4, which has high errors of measurement for 3 digit numbers as the sample sizes per case are of the order of 50, an indicator of progress is provided. The assessments can be observational, unobtrusive, and recorded in each case on the last recorded numeral said aloud, reported as a common vertical scale value.

Figure A8.4 Numbers in Estimated Order of Difficulty to Say Aloud-all numbers to 20, samples from thereon (Difficulties relative to '1')



Appendix 9 Chicago-Strategic Teaching and Evaluation of Progress (STEP)

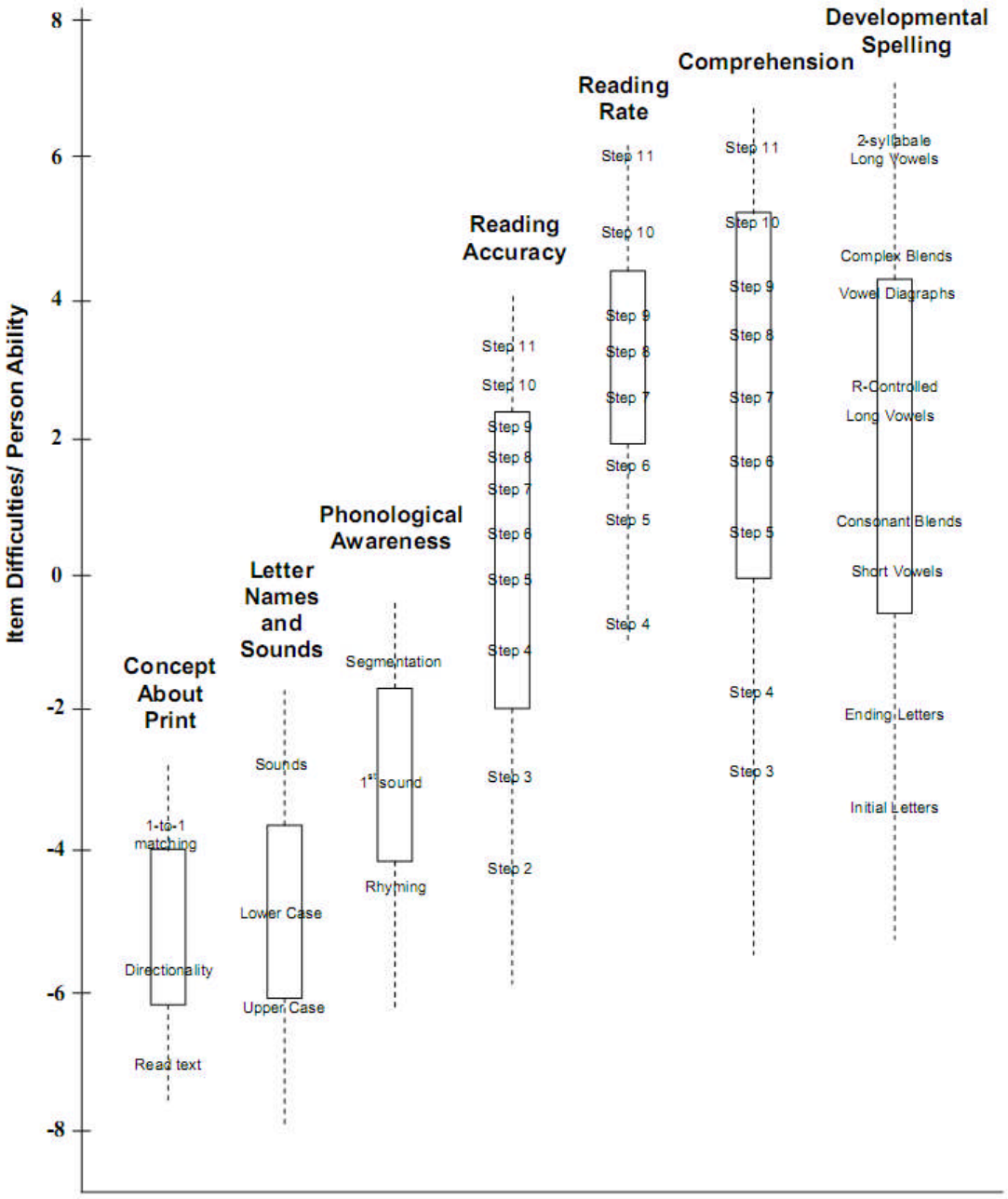
This case study develops a set of parallel scales of development to help monitor the learning progress in reading.

The Strategic Teaching and Evaluation of Progress (STEP) developmental assessment process for reading was developed at the Center for Urban School Improvement in Chicago over a ten-year period. The project worked with the Chicago Public Schools and others to study its impact and to test its performance by application of the Rasch model. STEP has much in common with other progress maps (DART, First Steps). The development process is documented by Kerbow & Bryk (2005) and is a key example in the annual 2008 *Brown Lecture in Education Research* of the American Educational Research Association by Raudenbush (2009).

The design of the system grew out of the Observation Survey from Reading Recovery (Clay, 1993) broadening a process applied in a tutoring context to the full classroom. The developers explored the psychometric properties of all the components of the survey and built up an appreciation of the difficulty relationships of each of the components. From this they were able to map out the relationship of the seven sub scales so that the scales, tasks within scales, and tasks across scales were aligned on a common logit scale. From this they were able to describe 12 steps from ages 5 to 8 (K-3 in their terms) that integrated the tasks from each of the subscales into a cohesive relationship of development. The process identified a general strategy for reading development over a 16 logit span from book orientation and letter recognition through to effective reading with fluency and comprehension at approximately 600 Lexiles and spelling words with double consonants 2-syllable vowel patterns by step 12.

The steps average about 1 logit apart but with the logit distance between steps decreasing (see Figure A9.1) as higher steps are achieved (consistent with other trajectories scaled in logits). The key factor determining improvement at the lower levels (steps 2 to 4) is explained as 'problem solving the words of the text'. Kerbow & Bryk believe students are 'learning enormous amounts about how letter patterns function and how to use this information to solve words' in these early steps. As students progress to higher text levels, the additional demand to reach decoding accuracy begins to decrease suggesting that the 'skill of problem-solving words and reading accurately becomes less of a hurdle as text levels increase.' (p. 51).

Figure A9.1 Overview of STEP subscales and step relationships (from Figure 3 Kerbow & Bryk, 2005)



While the step structure is used as the reference for 'data points' (that is as categories to indicate current reading status) the logit scale itself could be used as a complimentary progress scale for learning development, particularly as the logit scale better reflects an equality of increments than do the steps. As described above the spaces in logits between step achievement decreases with higher steps, making the 'step scale' increments unequal in 'difficulty'.

A key to the utility of the 'map' and aligned assessment processes is the 'visualisation' developed from the assessment information, which allows the teacher to document the development of each student in a form readily appreciated by the teacher, other teachers and the school generally. STEP reports provide clear intuitive graphical representations of student

status and growth (in steps). While initially developed as paper charts for each student, recorded as a 'wall chart', the concepts of the visualisation have been incorporated into computer screen reports generated by software into which progress records are easily entered. The record shows the current step for the student as well as the number of steps progressed since the beginning of the school year, that is both current status and rate of change, taken in readily by the clear colour assisted screen layout. The tool provides simple intuitive cues to remind teachers where each student is but the data are held in a form that can also be subjected to summary and general arithmetic for year-to-year and other analyses.

The visualisation concepts are not unique to the STEP project. Software has been developed to record student progress to support other developmental record schemes (Kidmap in Australia for Levels, First Steps or other ordered structures; mClass from Wireless Generation for DIBELS and PALS, Progress Assessment Series from Pearson) all include graphic reports.

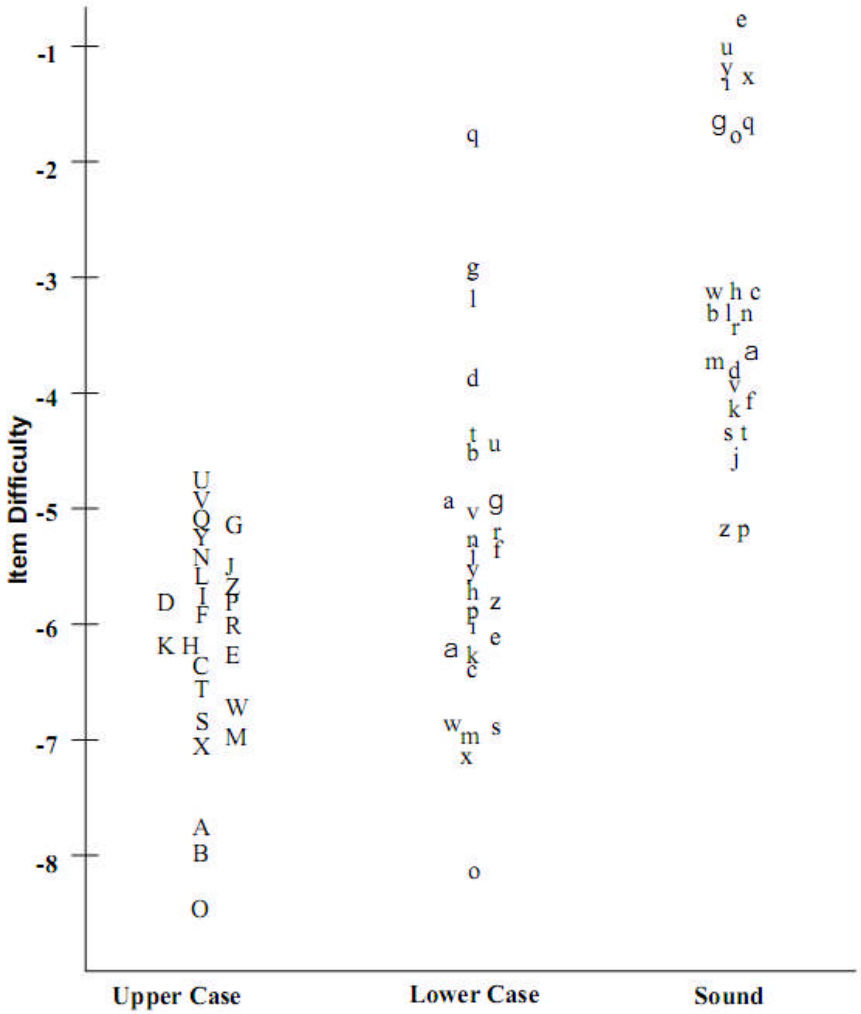
STEPs draws on the Lexile project to provide sets of levelled texts (difficulty determined using the Rasch calibrated Lexile analyser) that are the focus of individual conferences with the student (meeting the personalisation criteria of Fullan et al.). The teacher records reading accuracy and fluency, observes student reading behaviours, and engages students in comprehension conversations. Accompanying a given step are relevant assessment tasks that probe the skill profile on the other subscales to establish student status on these linked developments. The process has parallels with other probing conversation formats (NumPA from NZ as an example) and is acknowledged by the developers to be similar to other early literacy assessments but that 'the explicit combination of these tasks in developmental sequence is unique to the STEP Assessment and is organized on both a theoretical as well as an empirical basis.' (Kerbow & Bryk, 2005, p. 15.). Two parallel forms of the assessment at each step are provided (yellow and purple) to support repeat application of the assessment at a given step. The assumption is that two versions of the assessments will suffice since teachers do not use the assessment until their observations of the student confirm that the criteria for that step have been met. As a result the assessment is effectively confirming the teacher's observations. The second parallel version is used as a later follow-up assessment where a teacher over estimated the development of the student's reading.

Given the detail available in the developmental map, the step status makes clear the skills achieved and the skills to focus on for the next developmental period. STEPs is clearly well conceived, empirically developed and tested, and draws on the professional observational skills of the teacher. The step numbers provide a reference frame for student development (notwithstanding the non-equal increments). However, particularly in the beginning stages the step increments are large. The developers acknowledge as much:

> It should be noted that it may be possible to write additional texts that fall between Step 3 and Step 4 in difficulty. However, such fine-grained, formal assessment of text was not chosen because the information acquired is intended for classroom teachers. These smaller distinctions may prove very useful for one-on-one tutoring (such as Reading Recovery) but for thinking about instruction for small groups or whole classrooms such detail may be overwhelming. (Kerbow & Bryk, 2005, p. 51 footnote)

Overwhelming maybe, but perhaps also required somewhere in the teacher's kit if the knowledge of the within step progress status can assist the teacher decide what to do next. The almost 2 logit step increment offers the 'real estate' at least for considering some appropriately targeted early texts, although at this early stage 'reading' as such is very rudimentary. Figure A9.2 provides some insight into the development of pre-reading skills, similar to the number word development illustrated in the CEM example (Appendix 8). The items are the letters and sounds of the alphabet in difficulty order, with upper and lower case letters identified separately.

Figure A9.2 Overview of STEP Letter Identification and Letter Sound Item Maps (from Figure 5, Kerbow & Bryk, 2005)



Like the number map the letter map indicates the letters likely to be identified first. On the basis of the map O is the first letter and the first upper case letter identified. The lower case version is harder but is the second easiest letter to identify. B precedes A in upper case. The second lower case letter is x.

The range of difficulties is 7.0 logits from upper case O to lower case q. This is a wide range of difficulties, reflecting how hard it is for young children to master the alphabet as a prior step to reading. Confusing for the student is the differing order to vocalise the letters. Z and P are two thirds along the identification scale (that is relatively hard to identify) but the easiest to sound. The hardest to sound out is e, assumed because of the variety of sounds associated with the letter.

STEP provides an example of the combination of scale development and skill learning relationships on a common scale and the complexity of scale increment decisions. In particular the example illustrates the ways in which learning-task expected-development order can be established empirically. The empirically ordered items provide a framework for

structuring the learning, a framework for monitoring the learning and a scale with potential for adding finer resolution to the 'intra level zones' in levelled curricula. On the basis of the steps descriptions the upper end steps have a 0.52 logit increment between them (see Figure A9.1) while at the lower end this is 1.75 logits. It is clear from the STEP example, and the other examples in the trajectories chapter, that early learners move through large logit differences in skills development. The relative difficulties of the early learning elements (letters, numbers sounds) from easiest to the most difficult are wider than later Year level differences. Criteria made specific for teachers (based on the STEP type analysis) might increase the observation and assessment skills for these teachers providing them with a scale reference for documenting student learning development.

The general order of alphabet learning is corroborated by Justice, Pence, Bowles and Wiggins (2006). The correlation coefficient for the learning order for names of the 26 letters with the order in Figure A9.2 is 0.85 (n=339 students for Justice et al.).

Appendix 10 Individual learning trajectories.

The data analyses developed in Chapters 6 and 7 do not use longitudinal data for the observation of learning growth of individual students but snapshots at one point in time for a large set of individuals across 8 Year levels. As indicated in Chapters 5 and 6, the general trends in these cross sectional means of learning growth with Year level mostly approximate the general trend shown with the means of longitudinal data. Longitudinal data for individual students is another matter. Each trajectory is unique. As a consequence, teacher generated assessment data through observations will require different processes of recording and analysis than those conventionally applied in classrooms. This appendix considers some of the issues involved in individual student trajectories.

Whether teacher judgement data become an additional data source or not, schools will need to develop (or adopt) processes to better record and analyse student learning growth. There is pressure to do this generally and in particular from US policy analysts (Duncan, 2009; Rudner & Boston, 2003; Smith, 2008; State Educational Technology Directors Association, 2008). Unsurprisingly software and system suppliers (Ligon, 2009) also advocate the development of data warehouses and reporting process for individual student growth trajectories. The pressure is for better access to data and data driven decisions but "not advocating for additional high-stakes tests, instead … that using technology to assess students in a less formalized, yet more personalized, manner can glean benefits for teachers and students alike" (State Educational Technology Directors Association, 2008, p. 1). One source for this data will be computer adaptive testing. Another potential source is general classroom assessments, including teacher judgement assessments.

There is debate about how feasible it is to use summative, formative and interim assessments, to serve multiple purposes. Interim assessments are "assessments that fill the gap between classroom formative assessments and state summative assessments … an integral part of any comprehensive assessment system and should be evaluated as such" (Perie, Marion, & Gong, 2007, p. 1). However it is consistent with the concept of data warehousing, seen as underpinning the access to data and data mining, that as wide a range of assessment data as is possible is archived for a student. Were the data able to be stored using common scales across all assessments, it seems logical that these could generate a set of time related data points for each student.

As part of the thought experiment the nature of the data potentially available is considered. A data point could be stored with a minimum of five elements. These would be the student identifier, the strand of the curriculum, the source of the assessment, the assessment value and an automatic time stamp. Where data were not automatically recorded on the appropriate common scale, the source of the assessment might provide a conversion protocol to that scale. The student identifier provides the link to additional information about the student. Student and strand together provides a link to the class identifier. Entries could be automated (particularly if from other systems - adaptive testing, state tests etc.) or designed to minimise data entry requirements. One minimising control for teacher entries might be a policy of adding a new point for a student only when new development by the minimum scale unit has been observed. New wireless technologies have already been used to advantage to simplify teacher record keeping (Wireless Technologies, http://www.wirelessgeneration.com/).

Assuming 10 or so data points per year per student (per strand of the curriculum), how might this data be analysed within a year for the current teacher, and over a longer time scale (up to the whole school life of a student) for the benefit of a student and the other teachers with responsibility for that student? The question of when the student record would need to be destroyed is not discussed here but it is noted that the recording and/or the extended preservation of schools grades, as they would be seen, raises a significant 'privacy' issue.

If the trajectories of learning for individual students were to be made visible to teachers and students (through graphical presentation), what might the data and images look like? A visual representation of the data points for individual students showing the trajectory to the present position would provide a teacher with an understanding of the current status (a position with meaning; x=likely to be able to do this, unlikely to be able to do that) and the recent and previous rates of learning (the gradient with time from earlier points on the scale). This image in itself might be sufficient for a teacher. In principle, it might be feasible to enrich the teacher's understanding with estimates of likely progress points into the future. This appendix explores in general terms what individual trajectories of learning look like, based on available public data sources and the extent to which helpful forward projections of trajectories might be feasible.

Overview of the sources of longitudinal data for individuals

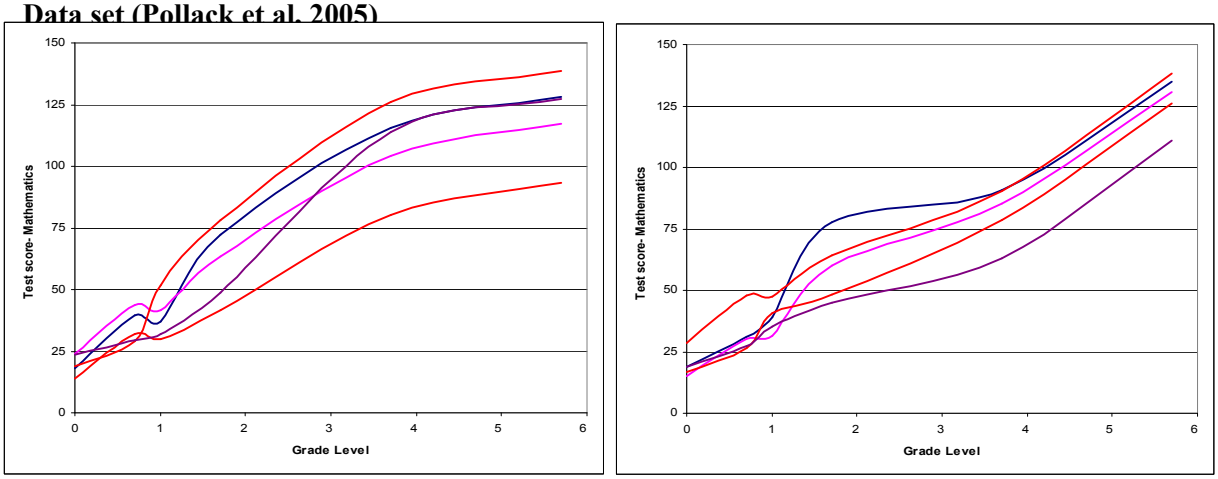
Data from the Early Childhood Longitudinal Study (ECLS) (Pollack, Atkins-Burnett, Najarian, &Rock, 2005) illustrate what individual trajectories look like for Year levels K to 5. Some analytical issues that might need to be incorporated into computer processes to support teachers with understanding and interpreting individual student trajectories are then considered. Examples of large longitudinal data sets are not readily found and many current initiatives that generate and manage individual student data seem to have moved into commercial products. As such, they are often protected from public access to data and processes. Contemporary approaches and analysis techniques for time recorded learning data built into commercial products were not found in the literature, although reviews of the products themselves are available (Ligon, 2009; What Works Clearinghouse, n.d.).

In the 1970s the issues being addressed in student trajectories in a then developing computer assisted learning project, were regularly published in the psychological literature. The early work of the Stanford University and Computer Curriculum Corporation (CCC) in the 1960s and 1970s, before it too became 'commercial in confidence', provides some understanding of the then thinking of how longitudinal records might assist in the management of learning.

Public examples of individual pathways

Figures A10.1 and A10.2 are panels of idealized learning trajectories taken from the US ECLS database. The trajectories are smoothed and are based on 6 points only (two in K, two in Grade 1, and single points in Grade 3 and Grade 5). Their purpose is to illustrate the wide variation in the pattern of pathways taken by students who start close together and may even and arrive at approximately the same point or vastly different points after 6 years. Figure A10.1 shows two sets of trajectories for Mathematics. The left panel illustrates students who start fast and rise to a score of 100 or above by Grade 3. The right panel shows a second group who have started slowly and then accelerated from Grade 3 to Grade 5. One case in the right panel shows a quick rise to a score of above 75 by Grade 1, no growth from there to Grade 3, and then high growth from Grade 3 to 5.

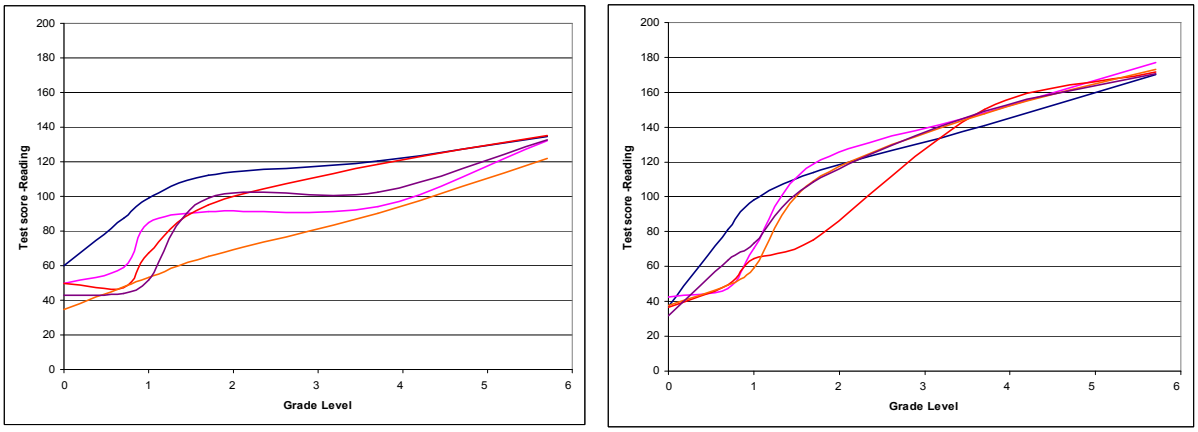
Figure A10.1 A sample of individual trajectories in Mathematics from the ECLS Public Data set (Pollack et al, 2005)



The panels in Figure A10.2 show another view of the destinations after 6 years, in this case for Reading. The left panel students start in a range from 38 to 60 and grow to 120. In the right panel the students start around the 40 score region, and via different pathways (similar to Figure A10.1) grow to around 170.

The purpose of the examples is to indicate the wide variability in trajectories and the complexity this raises for teachers in anticipating what might happen next and what intervention might be beneficial for each student. It is assumed a teacher would find a range of tools useful in dealing with this data, one part of the Fullan at al. Knowledge Base. Tools would include processes for estimating learning status (learning progressions), processes for visualising histories and plotting trajectories and indicators of what to do next at particular scale points. Examples that show the slow-fast-slow, steady or highly variable trajectories that students follow could be provided to help teachers identify outlier cases.

Figure A10.2 A sample of individual trajectories in Reading from the ECLS Public Data set (Pollack et al, 2005)



The data illustrated are very 'smoothed'. As data become available at shorter and more frequent intervals the complexity for teachers in making sense of the data will increase. This will be partly as a result of the increased variations around the 'true' trajectory due to measurement error. The reasons for the variations in trajectories not due to measurement error are beyond the scope of this thesis. Some recent research projects (Parrila, Aunola, Leskinen, Nurmi & Kirby, 2005; Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Hoeksma &

Kelderman, 2006) have applied Structural Equation Modelling, Hierarchical Modelling and Latent Class analysis to identify some of the contributing issues.

Making the trajectories visible

The visualisation of data is seen as a helpful process in its own right. At an earlier time in the apparently simple area of monitoring height, Burgess (1937) observed that many of the apparent irregularities of growth were really due to carelessness in measurement.

> The rate at which a child is growing is beginning to be regarded as one of the important indicators of his general physical condition. … Many school and private physicians who watch weight very carefully are content to measure height to the nearest inch often without regard to posture, or to measure with shoes sometimes on and sometimes off so that, according to their record; children apparently shoot up or shrink down in the most startling fashion. In many records, especially for younger children, heights and weights are transposed on the record card and much interpretation is needed to get the height picture approximately accurate. (Burgess, 1937, p. 305)

Burgess however makes the point, which the author believes will also apply when teachers have richer regular learning data (much from their own assessments and observation) and can see the trajectory of individual growth, that this will be its own incentive to develop better data procurement processes and to recognise where double checking will be required. What Burgess anticipated was that regular recording and graphing of height information would bring its own insights into error and irregularity, and alert observers to any anomalies.

> When a height chart is kept for the individual child, and his height line entered month by month or term by term an error in measurement stands out as dramatically as does a wrong thermometer reading on a fever chart. Where measurements are verified but the child suddenly begins to grow at an abnormal rate, either much faster or much more slowly than is usual, the graphic record gives parents, teachers and physicians prompt warning that he needs to be kept under close observation, and possibly given special medical care. The physician does not of course, make a diagnosis based on height alone, but a careful growth record is often a valuable diagnostic aid. (Burgess, 1937, p. 306)

The importance of height development per se may be less in the current context of generally better nutrition for children, but Burgess's insights about how serious readers of data react to anomaly are apt. Can a possible future where teachers react in the described manner to learning data be anticipated? Could individual development records on common scales help individual learning? A complicating issue for teachers is the uniqueness of each individual, trajectory, even though as illustrated in Chapter 5 the mean trajectories of groups of students can be modelled.

Each individual development trajectory is unique.

Since each individual trajectory is unique, to what extent can the patterns for groups, the average trajectory of the group with time, help in the understanding of individual trajectories? Keats (1983) cited by Willett & Sayer (1994), deemed models as having the property of *dynamic consistency* when the curve of the averages is identical to the average of the curves. Where dynamic consistency does not apply, the character of the individual curves is unrelated to the group trajectory, making it difficult to infer the shape of individual growth from a group growth curve. The variability in individual pathways illustrated in Figures A10.1 and A10.2 suggest that predicting an individual path at any time will be open to considerable uncertainty.

While increasingly sophisticated models for analysing change with time are available (Singer & Willett, 2003; Collins, 2006; Cudeck & Harring, 2007), models developed in this thesis for test data and teacher-assessed data are rather simple, particularly since the cross-sectional views are snapshots of time. It is impossible, for the author at least, not to wonder what the trajectories of students had been up to the point of the snapshot and where they might be at future points in time. This wondering raises the broad issue; can a 'black box' be developed for teachers to help them understand the implications of the pathways for each student and how to maximise their trajectories?

When processes for describing (modelling) individual student learning development over time within a class and across class years (grades) are considered, as might be required in such a 'black box' tool to help teachers with their decisions for appropriate types and timings of interventions, the feasibility of modelling individual trajectories needs to be addressed. How might this be done? What patterns might be expected? What might be the 'control boundaries' (in a quality control process) of development in say mathematics? When does a case change from being within expected ranges to being well outside? To what extent can group data assist with estimating 'safe' trajectories for individuals?

According to Molenaar (2004) "modern psychology is saturated with probability models and statistical techniques" (p. 202). However he believes that psychologists "attention is almost exclusively restricted to variation between individuals (interindividual variation [IEV]), to the neglect of time-dependent variation within a single participant's time series (intraindividual variation [IAV])" (Molenaar, 2004, p. 202).

He argues that most psychological processes should be considered as non-ergodic. The property of being non-ergodic implies a system that is influenced by history and is thus less predictable for lack of repetition of previous states. In contrast an ergodic system will return to states that are closely similar to previous ones.[37] Molenaar argues that the learning trajectory for an individual is non-ergodic. Furthermore, consistent with Keats's dynamic consistency, knowing the pattern with time of a population (IEV) does not necessarily assist in estimating the trajectory of an individual. In non-ergodic processes, an analysis of the structure of inter-individual variation will yield results that differ from results obtained in an analogous analysis of intra-individual variation.

> Hence for ... all developmental processes, learning processes, adaptive processes, and many more, explicit analyses of IAV for their own sakes are required to obtain valid results concerning individual development, learning performance, and so forth. (Molenaar, 2004, p. 202)

The essence on Molenaar's argument is that different approaches are required and different results are obtained when one follows individuals over time, as against aggregates of individuals. This point is made to support the complexity of the problem that faces teachers were more data provided to them, or developed by them, to follow the learning trajectories of individual students. Based on Molenaar's analysis any computer support system for the management of learning that depends upon simple extrapolations of individual trajectories from population patterns would be inaccurate. Recent publications (Molenaar & Campbell,

---

[37] Ergodic theory goes back to Boltzmann's ergodic hypothesis concerning the equality of the time mean and the space mean of molecules in a gas, i.e., the long term time average along a single trajectory should equal the average over all trajectories. The hypothesis was shown to be incorrect but the identification of a class of processes that have the property of tending to return to a previous state does provide a reference for considering 'converse' systems, particularly developmental systems.

2009; Molenaar, Sinclair, Rovine, Ram, & Corneal, 2009) indicate that Molenaar and colleagues believe there is still little literature and analytical support for non-ergodic cases:

> We are at the brink of a major reorientation in psychological methodology, in which the focus is on the variation characterizing time-dependent psychological processes occurring in the individual human subject. It will require substantial efforts from the community of psychological scientists to effectuate this reorientation. At present, there is very little literature on multivariate time-series designs and analysis techniques tailored to dealing with non-ergodic psychological processes. (Molenaar & Campbell, 2009, p. 116)

Molenaar et al. (2009) develop the argument for, and provide examples of, non-stationary time series modelling to address the problem of analysing individual level data.

> The EKFIS [extended Kalman Filter with iteration and smoothing] is a new and promising tool to analyse nonstationary time series in accordance with the classical ergodic theorems and with the basic tenets of DST [developmental systems theory]. Several aspects of the EKFIS are still under ongoing investigation, including alternative ways to determine the standard errors for the estimated time-varying parameters and technical aspects associated with the EM [expectation–maximization] loop in which the EKFIS is embedded as expectation step. Yet the results obtained thus far with the EKFIS indicate that it constitutes a viable and principled approach to the analysis of non-ergodic (nonstationary) developmental processes and thus allows for articulation of the basic tenets of DST—that individuals are complex dynamic systems, the characteristics of which are, themselves, changing and developing over time. (Molenaar et al., 2009, p. 369)

It is assumed that for data held on students from computer–adaptive testing, methods of analysis will be needed beyond the simple graphing of trajectories and summarising of norms especially where a reliable forward projection for an individual is expected. It is reassuring, for the author based on Molenaar's contemporary 2009 view, to appreciate that this problem is understood but not yet solved. The meagre literature search results appear to reflect that researchers are only at an early development stage for estimating trajectories for individuals. The provision of useful analytical tools and models of individual leaning growth, as required for teachers in the Fullan et al. Knowledge Base, will depend on further research.

Early attempts at individual based models of growth

There appear to be few sources for understanding individual trajectories as observed with small time increments. Of these few, a set of analyses come from the first major computer assisted learning projects, as a result of recording each student response. Starting in the 1960s a mathematical and practical approach to modelling of education development was explored and applied by Suppes and colleagues, based on work at Stanford University and the

Computer Curriculum Corporation (CCC)[38]. These models were developed to understand and predict trajectories for students involved in this early computer assisted instruction. In many cases the time dimension was defined by 'number of trials', a more precise measure than age or time on task. The materials were regarded as a behaviourist approach to learning development (Mazyck, 2002) but for the purposes of considering mathematical models of development, they provide example sets of individual trajectories and explorations of how they might be recorded and used for estimates of learning status at subsequent future time periods.

The two elements for generating a trajectory for learning for an individual are reliable measurement of learning and an adequate time metric. Where an IRT approach to measurement is applied, the measurement points on the time axis need to be spaced appropriately so that the change in learning over a unit of time is comparable to or less than the error of measurement for the learning status estimate. The frequencies of measurements, or the time intervals between measurements, affect the smoothness of the trajectory. Smoother, usually increasing trajectories of learning are generated when fewer measurements per unit time are applied (as shown in the smoothed example for ECLS in Figures A10.1 and A10.2 above). However, as each point is estimated with error, the more scale-readings the more likely the curve of best fit through the data points will reflect the 'true' trajectory. Most longitudinal studies cannot afford to measure at time increments below 6 months (for logistical reasons and because of the impact of testing in many cases on the students) and thus the number of points per individual is small, each measured with error. Data that show learning for individuals at small intervals between measurement points need less intrusive methods. Data derived from routine automated record keeping is one option for 'embedded measurement'.

This was achieved in the initial development of computer assisted learning assessments by recorded points being 'embedded' as a result of the mastery learning process. Measurement in this case is different to the IRT model. Problems of a specific type were repeatedly presented to the student, changing the values in the problem, until a specified criterion of consistent correct responses was met. An assessment data point was created when the criterion was achieved. While pedagogical processes have moved on from this stage, and the approach is now offered as a supplementary process only (SuccessMaker .http://www.pearsonschool.com/, June 2009), the early years of the process generated a unique opportunity to observe the shape of learning with time.

Suppes, Fletcher & Zanotti (1976) developed a simple set of 5 axioms to explore what they termed "student trajectories rather than student progress, in order to give the sense of a definite path as a function of time that we are predicting for the individual student." (p. 118).

The 5 axioms considered rates of processing, time effects, the effect of introducing new material, position (status) in a course (represented as a grade and progress in a grade in a decimal metric; e.g. 3.2 indicates grade 3, plus 0.2 of the way through grade 3 level) and the rate of progress in a course in relationship to the rate of introducing information in the course.

---

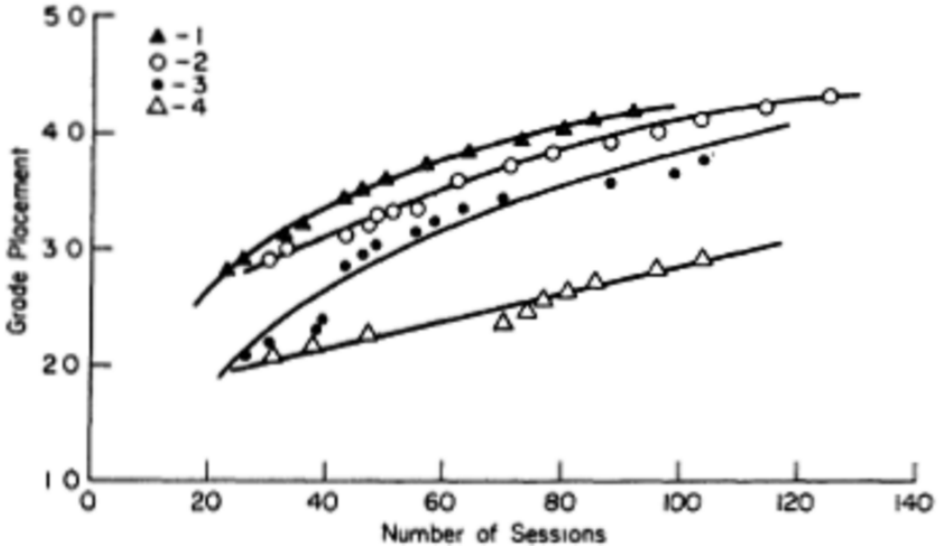
[38] Suppes founded the Computer Curriculum Corporation (CCC) in 1967. In 1990, it was acquired by Simon & Schuster, and then in 1999 by Pearson. One of CCC's major products was SuccessMaker. It is currently (2009) marketed by Pearson and includes 3,300 hours of supplemental instruction in English, language arts, math, science, and social studies in individualised, self-paced lessons, with the starting level individually determined and with diagnostic advice provided for recurring misunderstandings. **'**Forecasts tell you which students will meet instructional goals and when'. (Pearson website http://www.pearsonschool.com/, June 2009).

These considerations enabled Suppes et al.to develop some approaches to the general analysis of student trajectories in a maths curriculum covering roughly 7 years of elementary schooling. The curriculum was broken down into 14 parts, each corresponding to about half a year and included 14 strands (number concepts and decimals as two examples of strands) that were covered in some or all of the 14 time parts. Number concepts started in grade 1 and continued to grade 7.9, the only strand that occurred in all periods. Decimals, as an example, started in grade 4.0 and continued to grade placement 7.9.

A student's progress through the graded strands structure was a function of his/her own performance and was independent of the performance of other students. Progress on a given strand was also independent of performance on other strands. Movement through a strand used the pattern of correct and incorrect responses to insure a rate of movement that reflected performance. This structure has parallels with the general concepts of levels, strands and learning areas addressed in this thesis and described in Chapters 3 and 4. In particular, the individual progression of students meant that a student could be in a Grade 3 class but dealing with say grade 2.4 mathematics material, or for another student in the same class, material at grade 5.2.

Figure A10.3 below taken from Suppes et al.(1976), illustrates four typical cases from individual trajectories of almost 300 hearing-impaired students who participated in the program over a number of years. The grade placement value (GP) is the average of all strands for the student. The session number plotted was the one where the student had moved up (or down) 0.1 of a GP, achieved when students had worked through about 400 maths exercises, the actual number dependent upon error rates. This criterion for a plotted point ensures a relatively smooth curve as it reduces the error of measurement effect for points on the vertical scale that applies with IRT measurements.

Figure A10.3 Individual student trajectories- graphic from Suppes et al.(1976)



Three of the cases have lines of best fit that are clearly curves (1 to 3) while 4 is almost linear. Students 1 and 2 are very close after 20 sessions, but then follow different trajectories. Likewise for students 3 and 4 closeness after 20 sessions leads to quite different trajectories. Suppes et al.(1976) proposed that each trajectory could be described by the equation

$$y(t) = bt^k + c \qquad (1)$$

where y is the position (Grade Placement) at time t; b, c and k are parameters specific to the individual. While they describe the process as stochastic (as against deterministic) they do

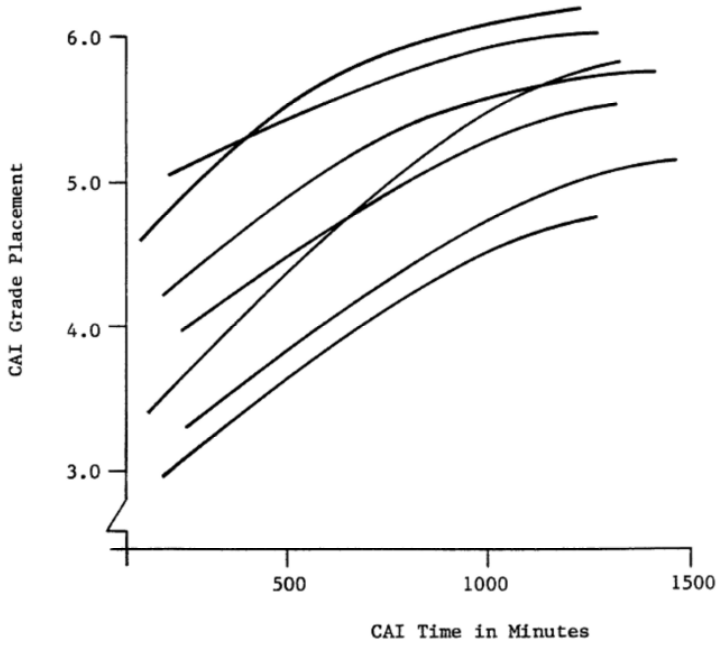
not include measurement error terms in the formulation (which is understandable given the basis for estimating GP discussed above).

The approach adopted was ground breaking in the 1960s and 70s and only a first exploration of the ways of modelling longitudinal learning data. Suppes et al.(1976) argued the need for a global theory of a student's progress through a curriculum (Suppes et al, 1976, p. 126).

Malone, Suppes, Macken, Zanotti & Kanerva (1979) developed 10 mathematical models based on 'power' functions of the general form of equation 1, for predicting a student's final grade placement. Data were obtained from 2000 elementary students at 2 weekly intervals for a full school year. Two simple models based on the most recent point and parameters estimated for the whole group, were the best for predicting end of year GP values. A power function model estimated individually for each student was best for describing all observed values for that student (using mean standard error over all students of data points compared to the fitted curves as the fit measure).

Later work in 1980 from the same general team (Macken, Suppes & Zanotti, 1980) argues against a global theory applied to the group without exploration of the patterns of learning over time (trajectories) for individuals. They presented Figure A10.4, making the point that while all the cases were one year below their chronological grade level each trajectory is distinct and a mean trajectory would not represent any one of the lines. This concern is similar to that raised more recently by Molenaar (Molenaar & Campbell, 2009).

Figure A10.4: Examples of individual student trajectories from Macken et al.(1980)



The individual trajectories can be described by the general equation (1) above only when the parameters b and c are estimated for the individual and not when they are estimated for the group. By implication, the estimates of the parameters for the individual are critical where judgements need to be made about whether an individual is performing outside the pattern that best describes their previous development.

A key insight is that the relationship between time and gain is not linear, even for individuals (or for grouped data as illustrated earlier). Macken et al. were concerned that evaluations of computer aided instruction (CAI) would misunderstand this point. Evaluations that assumed a linear relationship would mask some of the effects of individualised instruction. This thesis

argues that, at a broader level, assessments of any students over time that are not sensitive to the trajectory of the individual will misinterpret when and what assistance might be applied to individual students, even where more useful tools for teacher assessment are applied. As summarised by Macken et al, 'individuals proceed through the curriculum with distinct velocities and accelerations. The amount of gain per unit of time is different for different individuals and for the same individual across time.' (p. 82-83).

Conclusions about individual longitudinal records

As teachers are encouraged to connect records of learning for individuals over time they run the risk of being overwhelmed by the rich data they will have at hand. Without adequate analysis tools to make sense of the data, the benefits to students of better records of learning growth will not be obtained. Processes for managing these records must assume that the records come from wide range of sources (standardised and online tests, observations, class assessments, embedded assessments) and that a graphic history for each student can be displayed.

A range of analytical tools to help teachers understand their data can be anticipated. Issues that will be relevant in developing these tools include:

- Trajectories are idiosyncratic, and may not be able to be projected forward.
- Development of analytical models for individual development analysis is in its early days.
- Based on (possibly dated) research, and a possibly overly constrained learning process (CCC), it was necessary to estimate some individual parameters to project the likely pathways of development. Group data can be used to estimate some parameters only.
- Models based on the previously achieved point and previous estimates of rates of change are the most useful predictors of the next learning status point at t=x (implied in Molenaar et al. 2009 and Malone et al. 1979).
- At any point, the value of the learning scale has meaning in terms of what is says about a student's likely skills.

Appendix 11- Summary of equating approaches and issues

This appendix briefly summarises four approaches to equating. All four are used at various points in the analysis in Chapter 8. Traditional equating approaches include Mean, Linear and Equi-percentile equating (Kolen & Brennan, 2004; von Davier, Holland, & Thayer, 2004). A fourth process using independent Rasch scaling of the test and teacher scales and then equating the means and SDs, a Rasch scaled Linear equating, is also applied.

Mean equating

This equating process is the simplest, transposing the individual data points so that the mean of one scale equals the mean of the other. Generally the process is inadequate for effective transformation of scales, except in the rare case of equal SDs, as it takes no account of the difference in the spread of the two scales. It is mentioned briefly as it is used at the end of Chapter 8 in Figure 8.16 and subsequent figures to 'equate' scores to compare the model test data developed in Chapter 6 to the full Years 1 to 8 teacher data described in Chapter 7. It is used also as a process in this analysis to 'equate' the trajectories of the scales.

Linear equating

Linear equating transforms both the mean and the spread of one data set to match the other. The expression to do this is a simple linear transformation, and thus is insensitive to any non-linear relationship between the scales. Linear transformation is used in the Rasch equating described below with the teacher scale initially converted to equal interval units by the Rasch model analysis.

Equi-percentile equating.

The equi-percentile equating approach establishes the score at a number of percentile points on each scale and assumes that the scores from both scales at these points are equivalent. The relationship of one scale to the other need not be linear.

Rasch scaled linear equating

A Rasch model analysis of the full teacher assessment data set (not just the matched cases) is conducted and then the logit scores for the set of common students equated by the linear method. This brings the mean and SD of the set of teacher assessed students common to the test, to the test mean and SD. Test scores are not used as anchors for reasons discussed below

The transformation formula found for common persons is then applied over all teacher assessments to bring all teacher-assessed cases to a test score equivalent for all cases across all year levels. An estimate of model measurement error for each student is derived in the Rasch model analysis. This error from the profile level value to logit translation is one of the error factors additional to the possible variation in judgement skill and scale calibration of teachers.

Further issues in the equating process

Two additional issues arise in the consideration of the equating of the teacher and test scales: the timing of the assessments and the different levels of performance of a skill from a teacher or test perspective.

For the 1997 data there is an issue of timing. The students were tested in early August while teacher assessments extended over a period from early October to mid November. As a result the estimated learning status for students in October will, on average, be higher than at the point of testing. Based on the analyses in Chapter 7, where a relationship of learning with age based on teacher judgements is established, the real learning status at the earlier testing point would be about 0.1 profile units lower than that recorded by teachers. No direct adjustment is made for the time shift in this analysis. There is also a possibility that teacher judgement assessments were influenced by test results arriving before the teacher assessments were made. Based on the view taken by the teachers about test results and the lack of linking of the

test scale to the SPFAS scales, the likelihood of test results directly influencing teacher judgement assessments is very low.

Teacher judgement assessment data are treated, for the purposes of comparison, as if they occurred in August with a consequence that the equating arrangements will over estimate the relationship of profiles level to test scale units. Had the collection of data been repeated in subsequent years some form of adjustment would have been required. The data for 1998 were both collected in August removing the timing and influence problems.

The second issue that applies to both collection years is the potential difference in the criterion that teachers apply to a judgement of learning compared with the criterion in the test Rasch analysis model. In the Rasch model, as applied to the test data, a student is placed at a point where the odds of success on an item are estimated to be 50:50. The item scale itself is based on items positioned at the points where "the difficulty ... of an item … is the point on the latent variable (uni-dimensional continuum) at which the highest and lowest category have equal probability of being observed. For a dichotomous item, this is the point at which each category has a 50% probability of being observed" (Linacre, 2006, p. 300). This is the situation that applies for the test.

For the teacher judgement it is unlikely that a teacher will regard a level of performance of some behaviour as being achieved if the student can only perform it half of the time. Thus the equating process is for a teacher judgment at a higher criterion level than the test, for the same target behaviour. For the purpose of equating, this criterion shift, assuming it is relatively consistent across teacher assessments, will make no difference to general equating. It will however have implications in the interpretation of the relationship. Effectively the teacher scale will be displaced relative to the test scale, when performance of an actual skill is observed. Based on Masters et al. (1990) documenting the initial design of the Basic Skills Testing process for NSW, this concern is addressed in the conversion of students' test scores into Bands and the presentation of item difficulties in item maps, by rescaling the items to 0.7 probability of success rather than 0.5. However, when dealing with individual student data in Kidmaps (individual progress maps) the items are reported at their p=0.5 level. The data in this analysis are considered at the p=0.5 level. Based on the analysis in Chapter 6 the data analysed in this thesis were created at p=0.5.

As a result, taking the case of a specific behaviour, the test process will estimate it as 'achieved' well before the teacher. Since a focus on individual skills or behaviours is not considered directly in this analysis, this displacement will not effect teacher and test comparisons, which will be equated as if the difference between the scales does not exist. Further refinement of teacher judgements however would need to consider the practical implications. It also raises another source of variation in teacher judgement assessments. Teachers could all be generally aligned to the test scale, in principle, but apply idiosyncratic performance criteria, adding to the variability in aggregated teacher judgements.

Appendix 12 Summary of Rasch analysis statistics for teacher judgement assessment data

Table A12.1 1997 English

```
     SUMMARY OF 7868 MEASURED (NON-EXTREME) Persons
+--------------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE   ERROR       MNSQ   ZSTD   MNSQ   ZSTD  |
|--------------------------------------------------------------------------------|
| MEAN      80.0       2.9       -1.47     .26        .86    -.8    .86    -.8  |
| S.D.      33.2        .3        1.99     .08       1.54    1.7   1.54    1.7  |
| MAX.     208.0       3.0        3.13    1.13       9.90    9.9   9.90    9.9  |
| MIN.       1.0       1.0      -10.80     .05        .00   -3.9    .00   -3.9  |
|--------------------------------------------------------------------------------|
| REAL RMSE    .33  ADJ.SD    1.96  SEPARATION  6.00  Person RELIABILITY  .97  |
|MODEL RMSE    .27  ADJ.SD    1.97  SEPARATION  7.18  Person RELIABILITY  .98  |
| S.E. OF Person MEAN = .02                                                      |
+--------------------------------------------------------------------------------+
  MINIMUM EXTREME SCORE:      4 Persons
       VALID RESPONSES:  97.9%Person RAW SCORE-TO-MEASURE CORRELATION = .94
(approximate due to missing data)
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = 1.00 (approximate due to missing
data)

     SUMMARY OF 3 MEASURED (NON-EXTREME) Items
+--------------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE   ERROR       MNSQ   ZSTD   MNSQ   ZSTD  |
|--------------------------------------------------------------------------------|
| MEAN  209852.3    7705.0         .00     .00        .94   -1.9    .93   -2.5  |
| S.D.    1773.2      47.4         .02     .00        .19    8.2    .19    8.8  |
| MAX.  212360.0    7772.0         .01     .00       1.18    9.4   1.19    9.9  |
| MIN.  208583.0    7671.0        -.03     .00        .72   -9.9    .73   -9.9  |
|--------------------------------------------------------------------------------|
| REAL RMSE    .00  ADJ.SD     .02  SEPARATION  3.47  Item   RELIABILITY  .92  |
|MODEL RMSE    .00  ADJ.SD     .02  SEPARATION  3.58  Item   RELIABILITY  .93  |
| S.E. OF Item MEAN = .01                                                        |
+--------------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00 (approximate due to missing data)
23115 DATA POINTS. APPROXIMATE LOG-LIKELIHOOD CHI-SQUARE: 100312.54
```

No of iterations = 741

Figure A12.1 1997 English-Distribution of Infit Mean Square values of fit to Rasch model.

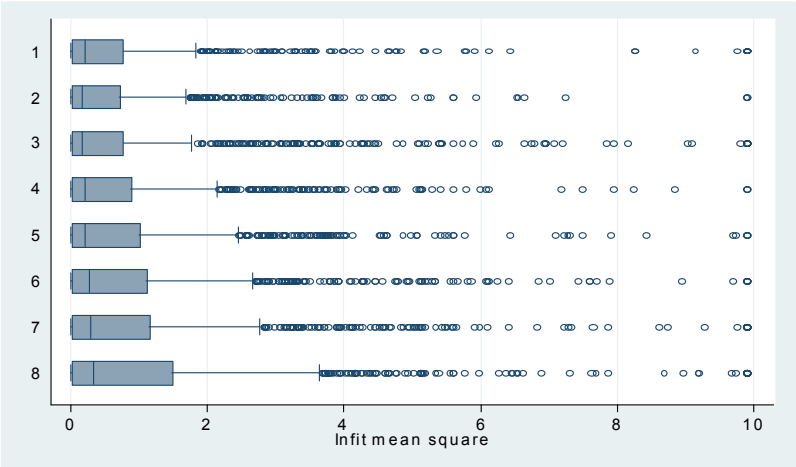


Table A12.2 1997 English subset Years 3 and 5 with both Test and Teacher assessment n=1275- Means and SDs equated for the common subset and then solution applied to all cases.

		Test scores	Teacher Rasch Analysis No Anchor	Teacher Rasch Mean and SDs transformed to match test	Teacher Rasch No anchor Original Measurement error	Teacher Rasch No anchor re-scaled Measurement error
Matched Y3 & Y5 only	Mean	1.06	-1.64	1.06	0.26	0.26
	SD	1.37	1.33	1.37	0.06	0.06
	N	1275	1275	1275	1275	1275
All cases	Mean	1.06	-1.47	1.23	0.26	0.27
	SD	1.37	2.00	2.07	0.09	0.09
	N	1275	7872	7872	7872	7872

Table A12.3 All Yrs 1997 Profiles Assessments: Largest Standardized Residual Correlations Used To Identify Dependent Items

RESIDUL CORRELN	ENTRY NUMBER Item	ENTRY NUMBER Item
-.65	2 Writing	3 SpeakListen
-.56	1 Reading	3 SpeakListen
-.27	1 Reading	2 Writing

Table A12.4 All Yrs 1997 Profiles Assessments Table Of Standardized Residual Variance

	Empirical		Modeled	
Total variance in observations =	96.8	100.0%		100.0%
Variance explained by measures =	93.8	96.9%		96.7%
Unexplained variance (total) =	3.0	3.1%	100.0%	3.3%
Unexplned variance in 1st contrast =	1.7	1.8%	57.9%	
Unexplned variance in 2nd contrast =	1.3	1.3%	42.1%	
Unexplned variance in 3rd contrast =	.0	.0%	.0%	
Unexplned variance in 4th contrast =	.0	.0%	.0%	
Unexplned variance in 5th contrast =	.0	.0%	.0%	

Table A12.5 1998 Mathematics

```
     SUMMARY OF 12139 MEASURED Students
+-------------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-------------------------------------------------------------------------------|
| MEAN     104.0       3.9        -.73     .19       .71   -1.0    .71   -1.0  |
| S.D.      46.6        .4        1.59     .06      1.19    1.6   1.20    1.6  |
| MAX.     270.0       4.0        3.93    1.12      9.90    9.9   9.90    9.9  |
| MIN.       1.0       1.0       -8.60     .13       .00   -3.9    .00   -3.8  |
|-------------------------------------------------------------------------------|
| REAL RMSE    .23  ADJ.SD    1.57  SEPARATION  6.82  Studen RELIABILITY  .98  |
|MODEL RMSE    .20  ADJ.SD    1.57  SEPARATION  7.81  Studen RELIABILITY  .98  |
| S.E. OF Student MEAN = .01                                                    |
+-------------------------------------------------------------------------------+
      LACKING RESPONSES:      8 Students
        VALID RESPONSES:  97.9%
Student RAW SCORE-TO-MEASURE CORRELATION = .95 (approximate due to missing data)
CRONBACH ALPHA (KR-20) Student RAW SCORE RELIABILITY = .98 (approximate due to missing
data)

     SUMMARY OF 4 MEASURED Profile levels
+-------------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-------------------------------------------------------------------------------|
| MEAN  315703.0   11889.7         .00     .00       .75   -9.9    .73   -9.9  |
| S.D.   14358.9     211.0         .08     .00       .10     .0    .10     .0  |
| MAX.   332136.0  12104.0         .11     .00       .87   -9.9    .85   -9.9  |
| MIN.   293402.0  11543.0        -.10     .00       .59   -9.9    .58   -9.9  |
|-------------------------------------------------------------------------------|
| REAL RMSE    .00  ADJ.SD     .08  SEPARATION 24.57  Profil RELIABILITY 1.00  |
|MODEL RMSE    .00  ADJ.SD     .08  SEPARATION 24.57  Profil RELIABILITY 1.00  |
| S.E. OF Profile leve MEAN = .05                                               |
+-------------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000
Profile level RAW SCORE-TO-MEASURE CORRELATION = -.99 (approximate due to missing
data)
47559 DATA POINTS. APPROXIMATE LOG-LIKELIHOOD CHI-SQUARE: 214755.62
```

No of iterations = 275

Note: 89 cases subsequently deleted due to Teacher assessments providing zero data or only one of four strands (items).

Figure A12.2 1998 Mathematics-Distribution of Infit Mean Square values of fit to Rasch model

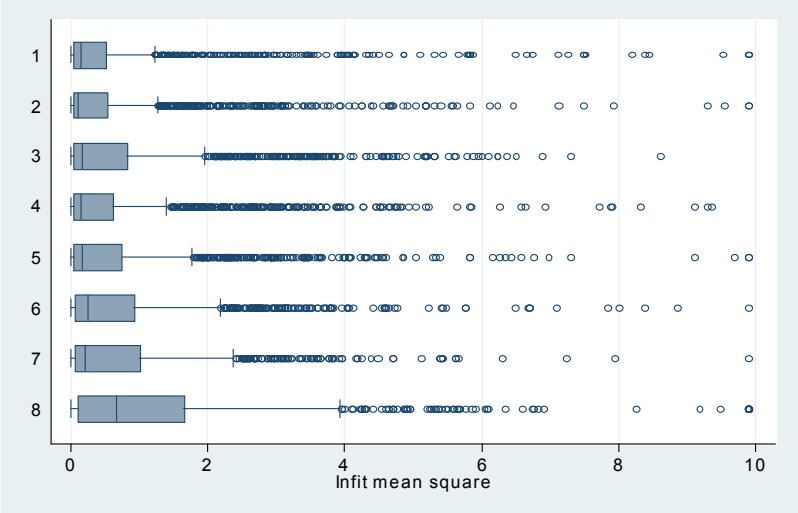


Table A12.6 1998 Mathematics subset Years 3 and 5 with both Test and Teacher assessment n=2105- Means and SDs equated for the common subset and then the solution applied to all cases

		Test scores	Teacher Rasch No Anchor	Teacher Rasch Mean and SDs transformed to match test	Teacher Rasch No anchor Original Measurement error	Teacher Rasch No anchor Re-scaled Measurement error
Matched Y3 & Y5 only	Mean	0.77	-0.70	0.76	0.18	0.27
	SD	1.44	0.99	1.44	0.03	0.04
	N	2105	2105	2105	2105	2105
All cases	Mean	0.77	-0.72	0.74	0.19	0.28
	SD	1.44	1.58	2.28	0.06	0.09
	N	2105	12050	12050	12050	12050

Table A12.7 All Yrs 1998 Profiles Assessments: Largest Standardized Residual Correlations Used To Identify Dependent Items

RESIDUL CORRELN	ENTRY NUMBER Profile lev	ENTRY NUMBER Profile lev
-.47	1 Chance	3 Number
-.37	3 Number	4 Space
-.33	2 Measurement	4 Space
-.30	1 Chance	2 Measurement
-.28	1 Chance	4 Space
-.23	2 Measurement	3 Number

Table A12.8 All Yrs 1998 Profiles Assessments: Table Of Standardized Residual Variance

		Empirical		Modeled	
Total variance in observations	=	155.6	100.0%		100.0%
Variance explained by measures	=	151.6	97.4%		96.5%
Unexplained variance (total)	=	4.0	2.6%	100.0%	3.5%
Unexplned variance in 1st contrast	=	1.5	1.0%	37.8%	
Unexplned variance in 2nd contrast	=	1.3	.8%	32.7%	
Unexplned variance in 3rd contrast	=	1.2	.8%	29.5%	
Unexplned variance in 4th contrast	=	.0	.0%	.0%	

Figure A12.3 Comparisons of 'Item' difficulty relationships

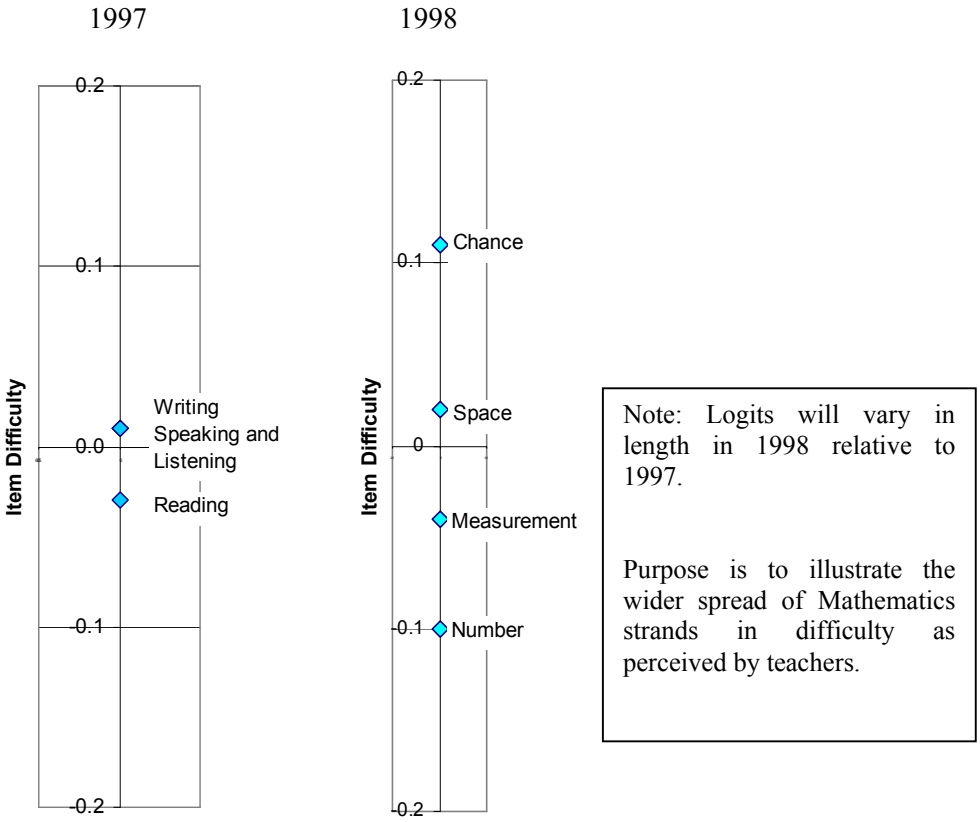
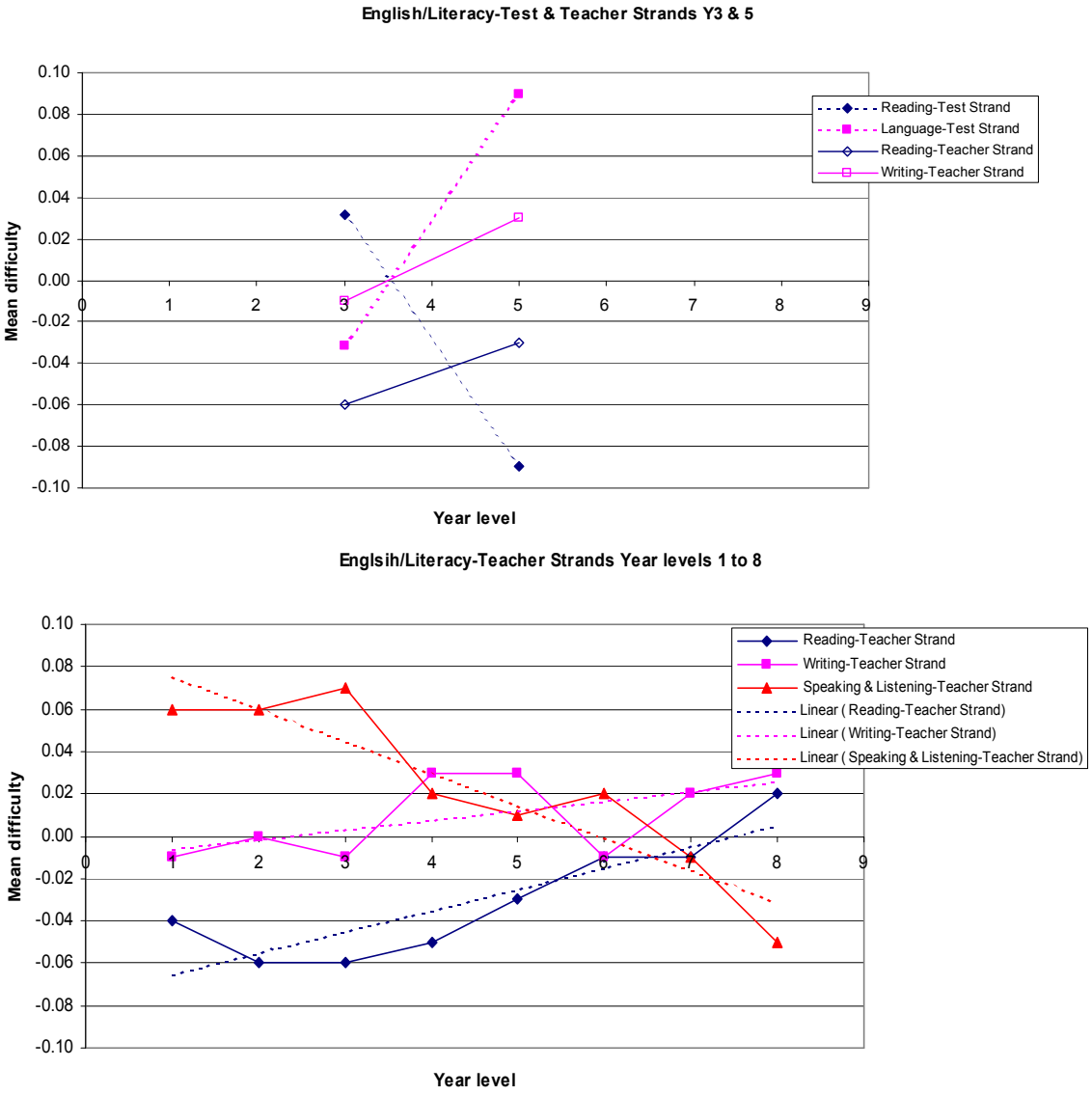


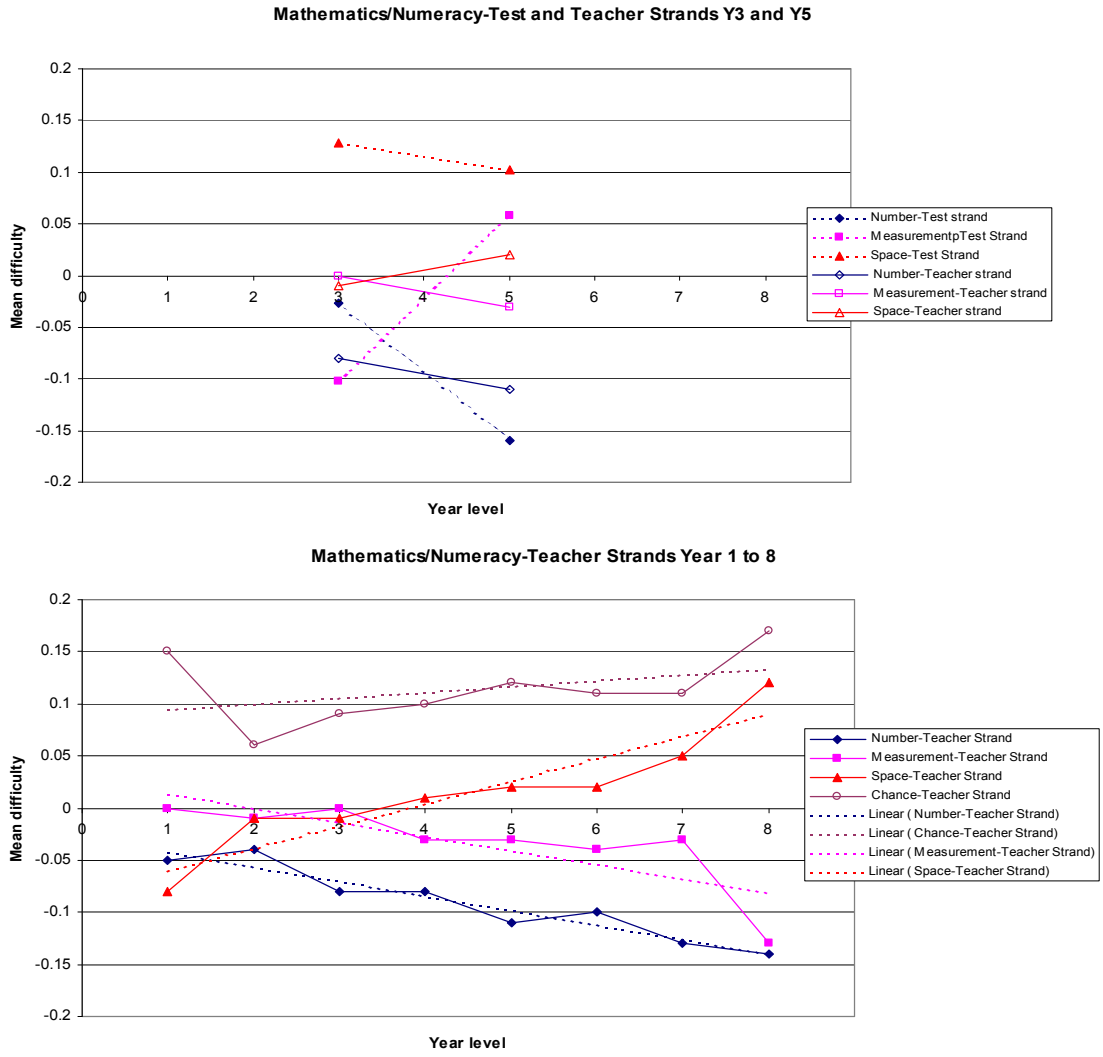
Figure A12.4 English Teacher and Test assessments compared by Strand difficulty



The top panel compares Year 3 and Year 5, the only Year levels for which data from both sources occurred. Test strand data were created from test population means of sets of items designated as Reading or Language. Language includes elements of writing, spelling and grammar and is most likely not directly comparable to Writing. Strand difficulties are created by the difference between population means with the mid point in difference set to 0 and the general scale reversed so that 'easy' is lower on the scale. On this basis Test Strand difficulties were about 0.05 logits apart at Year 3 with Language easier. By Year 5 they were 0.18 logits apart and with Reading now easier than Language. For Teacher assessments Reading was easier than Writing and while both became more difficult they stayed in the same general relationship.

The lower panel shows the trends in strand difficulty by Year level based on Differential Item Function. (In this analysis strands are items.) As Year level increases Reading and Writing as seen by teachers appears to get harder; Speaking and Listening becomes easier.

Figure A12.5 Mathematics Teacher and Test assessments compared by Strand difficulty



The top panel compares Year 3 and Year 5, the only Year levels for which data from both sources occurred. Test strand data were created from test population means of sets of items designated as Number, Measurement and Space (Chance is not identified in the Test design). Strand difficulties are created by the difference between population means with the mid point in difference set to 0 and the general scale reversed so that 'easy' is lower on the scale.

On this basis Test Strand difficulties were about 0.22 logits apart at Year 3 with Measurement the easiest. By Year 5 the spread has become 0.26 logits apart and with Number now easier than Measurement. Space is hardest in both periods. For Teacher assessments Number was easier than either Measurement or Space. Number and Measurement become less difficult by Year 5. Space is consistently the hardest in both assessment processes.

The lower panel shows the trends in strand difficulty by Year level based on Differential Item Function. (In this analysis strands are items.) As Year level increases Number and Measurement as seen by teachers appears to get easier; Space and Chance become harder and Chance remains the strand seen as hardest to achieve a high assessment.

Appendix 13. Estimates of the proportions of teachers at various levels of correlation and match to the test scales.

Table A13.1 tabulates the correlation coefficients for the majority of sites with students assessed by both teachers and tests for 1997 and 1998. The combined data include an estimated 600 teachers. About 120 teachers/sites are excluded on the basis of small number of students (<5) due to the relatively higher correlations that these very small sets generate. The tabulation is a very broad estimate only of the proportions of teachers in each cell.

Table A13.1 Estimates of the percentage of teachers in categories of correlation with the tests and rate of match to the test

Correlation Coefficient Category	Match Rate 0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	Total	Accum-ulated %
1			0.1%	0.4%	0.1%	0.1%		1.1%			1.2%	3.2%	3.2%
0.9	0.3%		0.6%	0.4%		2.4%	2.8%	2.5%	1.2%	0.7%	0.4%	11.2%	14.4%
0.8	0.1%	0.3%	1.7%	0.8%	1.1%	1.9%	8.2%	4.7%	5.1%	0.7%		24.6%	39.0%
0.7		0.3%		1.1%	2.8%	6.4%	4.3%	4.1%	1.0%	1.1%		21.0%	60.0%
0.6		1.1%		0.1%	2.6%	8.3%	6.1%	2.1%	2.1%		0.1%	22.5%	82.6%
0.5	0.1%	0.3%	0.1%		0.6%	1.2%	3.0%	1.1%	0.8%			7.3%	89.9%
0.4				1.1%	2.5%	0.7%	0.6%	1.2%				6.1%	96.0%
0.3				0.4%		0.4%		0.1%				1.0%	97.0%
0.2						1.1%						1.1%	98.1%
0.1			0.1%			0.4%		0.3%				0.8%	98.9%
0	0.1%											0.1%	99.0%
-0.1												0.0%	99.0%
-0.2				0.1%								0.1%	99.2%
-0.3								0.1%				0.1%	99.3%
-0.4				0.1%								0.1%	99.4%
-0.5				0.1%								0.1%	99.6%
-0.6												0.0%	99.6%
-0.7				0.3%								0.3%	99.9%
-0.8								0.1%				0.1%	100.0%
Total	0.7%	1.9%	2.6%	5.1%	9.7%	23%	25%	18%	10%	2.5%	1.8%	100.0%	
High Correlation Low Match	0.4%	0.6%	2.4%	2.8%	4.0%							10.1%	
Low Correlation Low Match	0.3%	1.4%	0.3%	2.4%	5.7%							10.0%	
High Correlation High Match						11%	15 %	12 %	7.3%	2.5%	1.7%	49.9%	
Low Correlation High Match						12%	9.7%	5.1%	2.9%	0.0%	0.1%	30.0%	

Estimates are of the proportions of teachers providing assessments in each of the categories of match. The data are a combination of the 1997 and 1998 cases. The correlation coefficients for each broad category are cross tabulated with the match rates to the test, that is the site specific match rate and correlation are weighted by the estimated number of teachers at the site. The site performance is ascribed to all the teachers estimated to be at the site. As a result the estimates diminish the variability in teacher performance.

Aggregating the data for both 1997 and 1998 hides some minor differences in the distribution of the matches and correlations by learning area. The values for correlations and match rates

are categorised by rounding to one decimal point, effectively centring the points on the listed values.

The table is segmented into four sectors on the basis of arbitrary definitions of High and Low correlation and High and Low match. High correlations are defined as those above 0.6 (about 60% of teachers), High matches are those 0.5 and above (about 80% of teachers). Cross checking shows that the mean match rate for all the teachers is 0.57, slightly higher than the expected match rate overall of about 0.54 established in Chapter 8. The mean match rate of the teachers with a 'high' match rate (estimated to be 80% of teachers) is 0.64. The mean match rate for the remaining 20% of teachers is 0.3. Combined for all teachers this can be shown to average out at 0.57 as above.

The tabulation indicates that the overall match rate is obtained from a wider range of teachers than the simple assumption that some teachers match for all assessments and some match for none. The relatively large proportion of teachers with greater than 0.5 match rates implies that those sites/teachers excluded due to small numbers of cases also had lower than average match rates.

An estimated ninety six percent of teachers have correlations with the test at 0.4 or higher. This highlights that teachers on the whole, order students in general terms in the same broad order as the test. When they do this they may not meet the criteria for a match. On the High/Low criteria in the table, 10% of teachers have 'high' correlations but low matches. This implies these 10% are displaced from the test scores in such a way that their assessments lay outside the control lines.