# 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.