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.