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

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

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

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

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

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

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

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

Issues in comparing teacher judgement and test assessments

Assessment resolution for tests and teachers

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

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

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

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

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

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

Variability in assessment skill and calibration within and between teachers

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

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

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

Teacher judgement reliability

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

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

Teacher preparation for assessment

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

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

Lack of research

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

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

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

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

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

Methods comparison

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

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

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

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

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

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

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

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

Table 4.1 Table of Kappa values

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

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

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

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

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

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

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

Value of the scatter plot techniques

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

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

Forms of match between independent assessments

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

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

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

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

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

Clarifying how teachers make judgement assessments.

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

---

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

US research

Two different interests affect the focus of US research.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

US research conclusions

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

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

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

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

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

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

Teacher Assessment in England

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

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

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

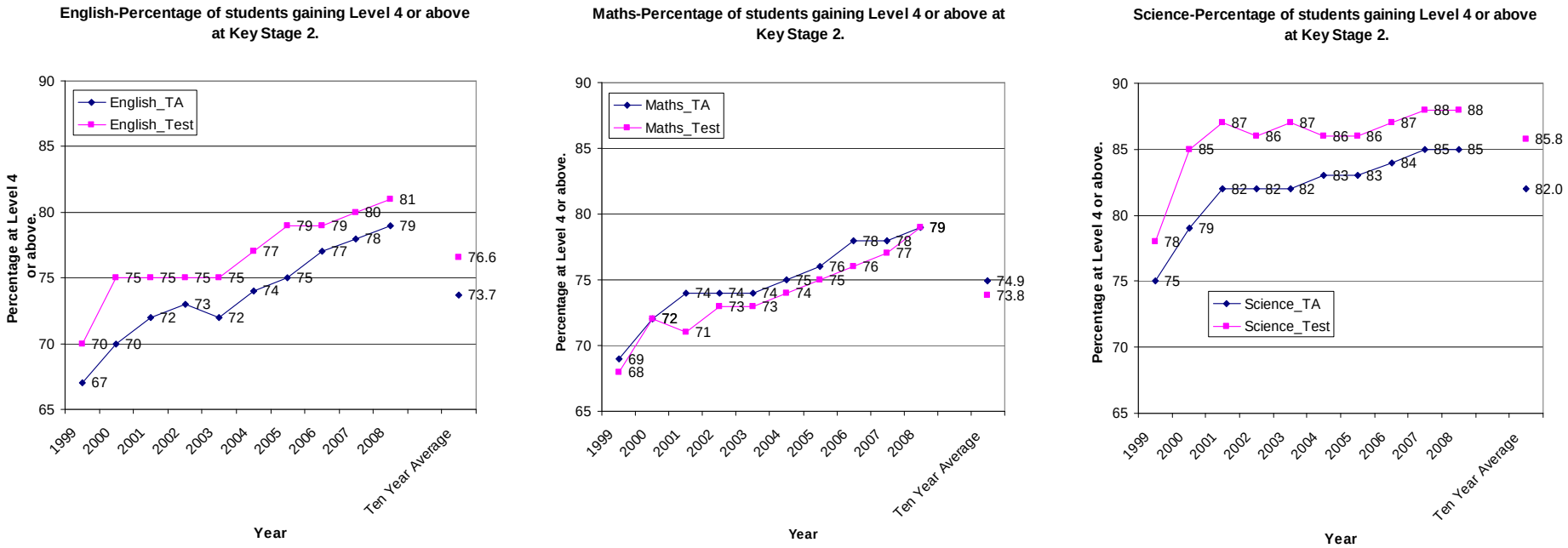
However being aware of the patterns of the two assessment processes over time and the persistence of these patterns by subject provides some hints as to their relationship.

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

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

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

Figure 4.1 Time Series of Teacher Assessments (TA) compared with Test Assessments. Percentage achieving at or above Level 4 for 11 year olds (Key Stage 2)- England



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

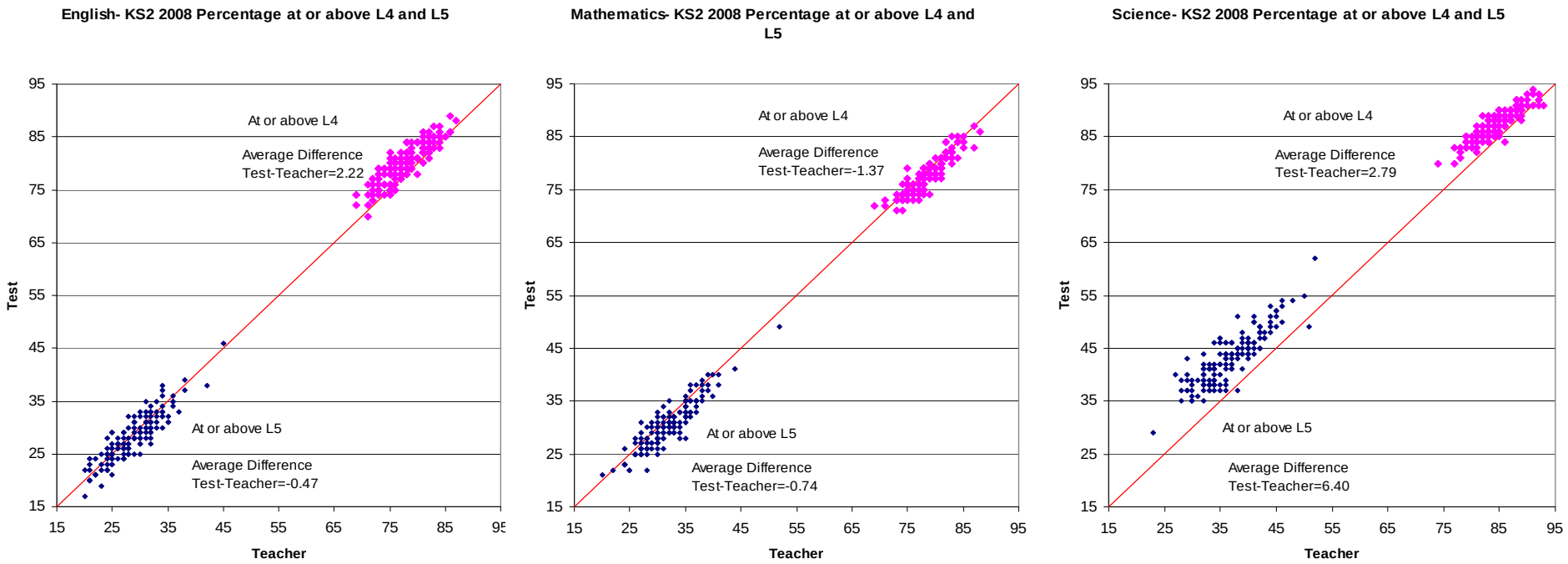
1999-SFR29/1999, 2000-SFR43/2000, 2001-SFR37/2001, 2002-SFR 21/2002, 2003-SFR 20/2003, 2004-SFR 30/2004, 2005-SFR 31/2005, 2006-SFR 31/2006, 2007-SFR 24/2007, 2008-SFR19/2008, SFR 06/2009,

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

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

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

Figure 4.2 Teacher Assessments (TA) compared with Test Assessments. (2008), by Local Authority (LA). Percentages achieving at or above Level 4 and Level 5 for 11 year olds (Key Stage 2) England



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

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

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

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

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

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

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

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

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

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

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

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

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

Table 4.2 Summary of Matches of Teacher and Test Assessments -Worcestershire LEA; data for 1997 to 2001 combined, with level as unit of reporting.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Summary of Teacher Judgement in England

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

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

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

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

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

Approaches to teacher judgement in Australia

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

Case 1- National Assessment –Language and Literacy

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

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

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

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

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

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

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

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

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

Case 2- Data from the Victorian Auditor-General 2009

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

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

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

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

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

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

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

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

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

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

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

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

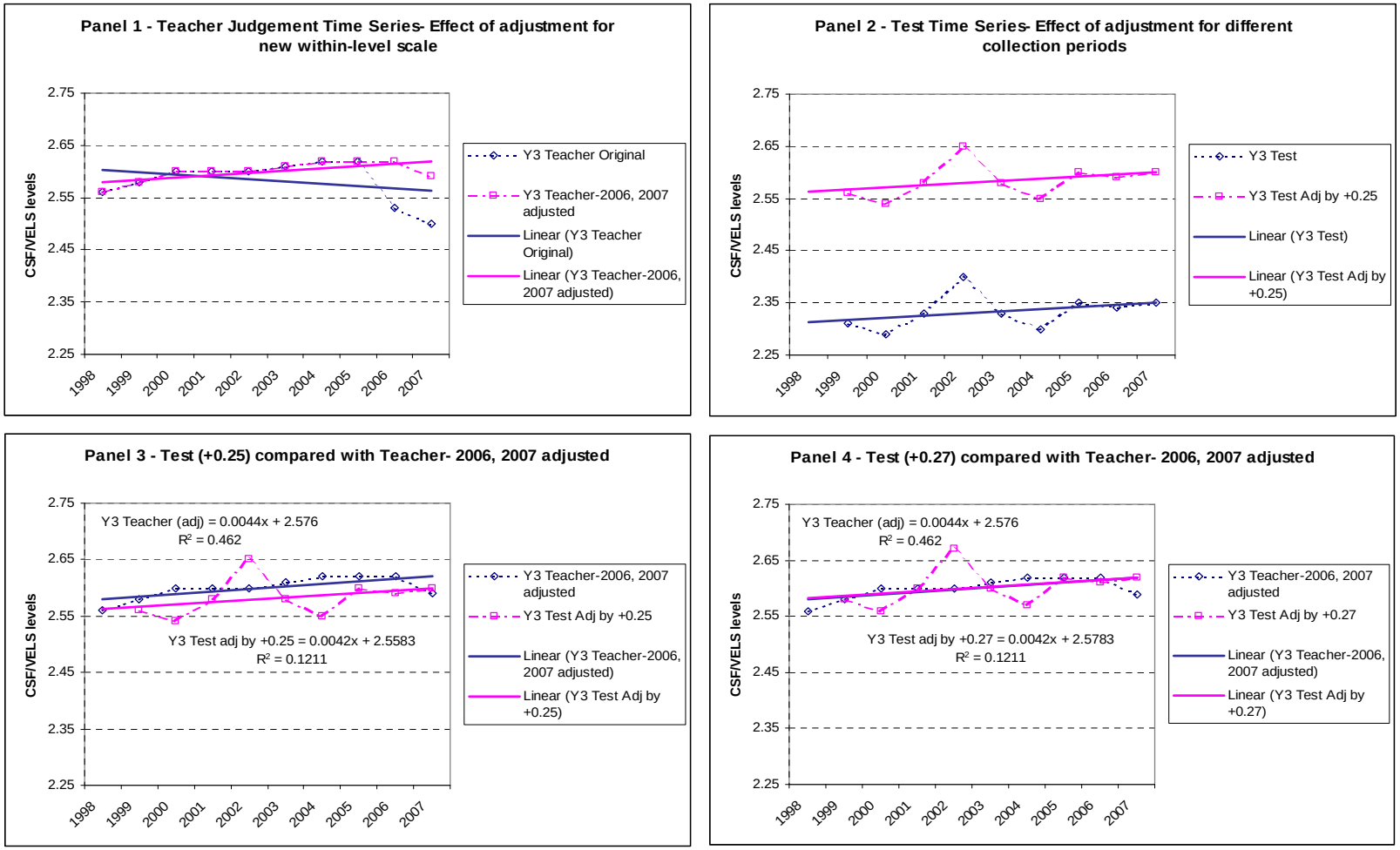
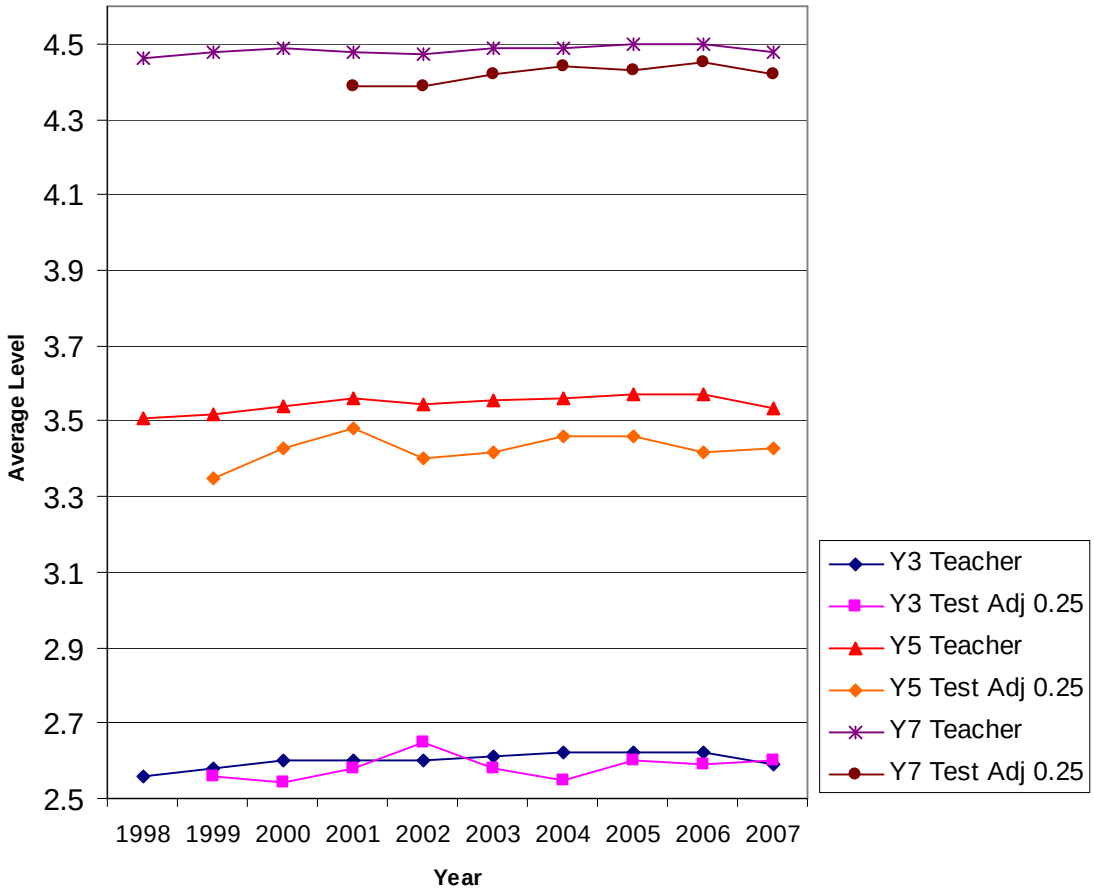


Figure 4.4 Comparison of Times series of Year 3, 5 and 7 Teacher and Test Data –Reading (values estimated from original graphs in Victorian Auditor-General, 2009)



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

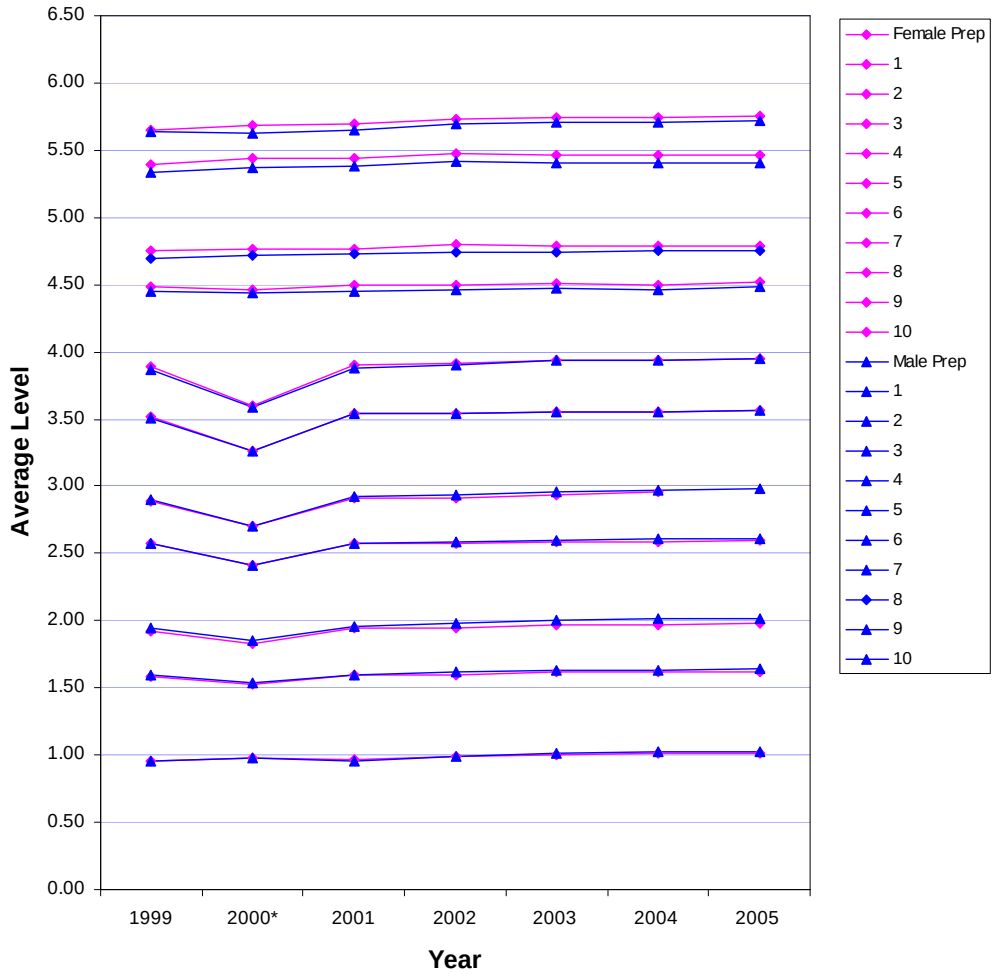
The two assessment processes, allowing for the general errors of measurement and estimation, produce very similar results at a population level. Where there are differences between the two sources of assessment, the differences are very consistent suggesting that there are possibilities for recalibrating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

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

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

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

Figure 4.5 Comparison of Times series of Years P-10 Teacher Assessment Data –Mathematics (Number 1-6/Chance and Data 7-10), by gender



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

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

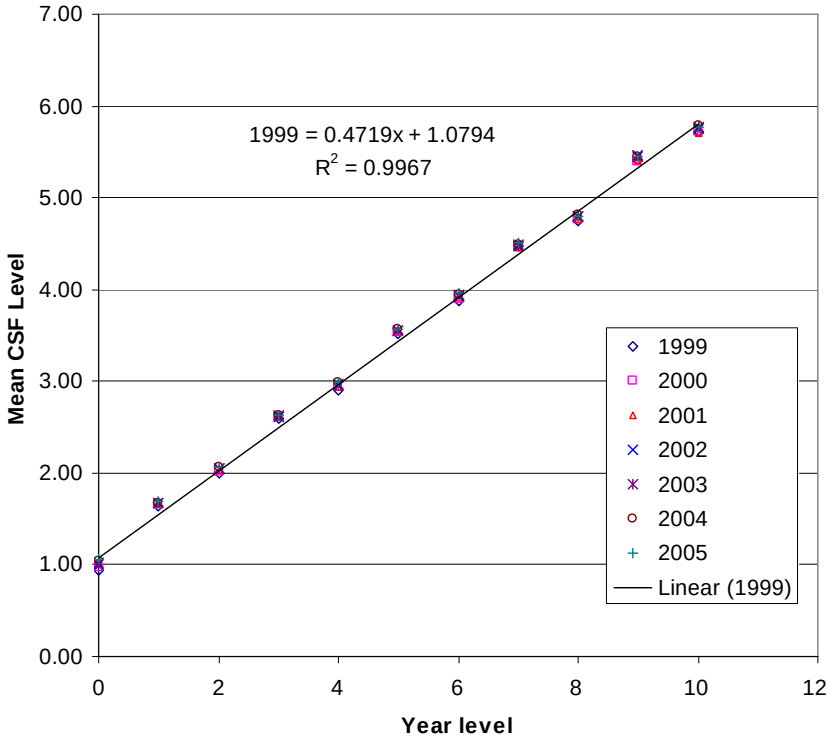
5, 7 & 9) are slightly higher on average than where they would be if the annual growth per Year level were even. The values in the tested Year levels may be affected by the moderation effects of the feedback of test results to teachers in these Year levels.

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

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

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

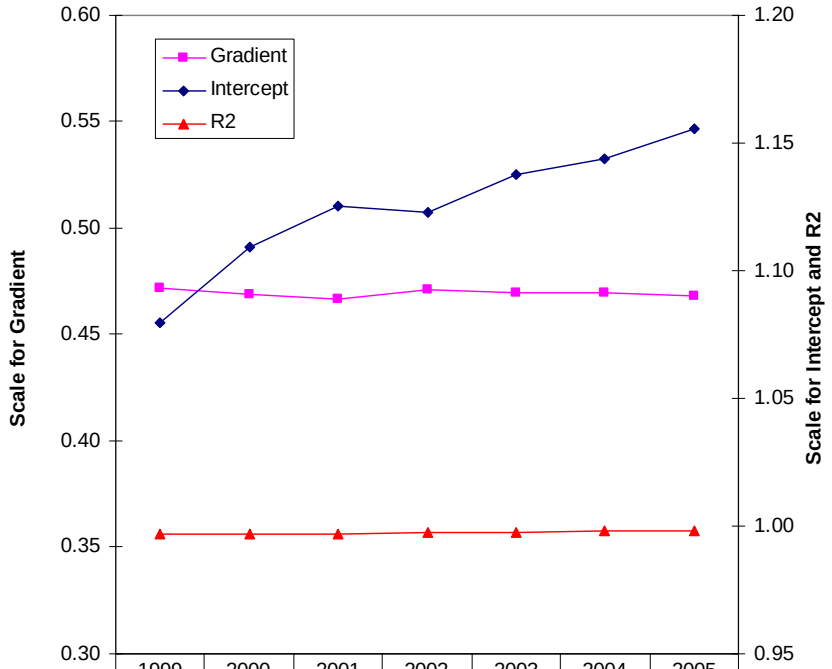
Figure 4.6 Reading: All students-Mean Teacher Judgement Assessments 1999-2005 by Year level.



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

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

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



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

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

Case 3- Quality Schools Project

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

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

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

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

---

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

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

Case 4- Online structured assessment interviews that require teacher judgement

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

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

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

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

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

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

Case 5-Tasmania -Indirect evidence- validation studies

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

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

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

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

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

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

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

Cases 6 & 7-Queensland and the Australian Capital Territory

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

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

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

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

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

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

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

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

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

Case 8- South Australia

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

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

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

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

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

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

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

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

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

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

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

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

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

Do accurate teacher assessments influence learning?

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

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

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

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

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

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

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

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

A summative overview of teacher judgement -Harlen

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

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

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

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

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

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

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

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

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

Summary

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

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

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

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

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

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

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

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

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

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

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

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