UNIVERSITY OF
OXFORD

OXFORD UNIVERSITY
CENTRE FOR EDUCATIONAL ASSESSMENT

REVIEW OF TEACHER ASSESSMENT: EVIDENCE OF WHAT WORKS BEST AND ISSUES FOR DEVELOPMENT

Gordon Stanley, Robert MacCann, John Gardner, Laura Reynolds and Imogen Wild

March 2009

QCA Qualifications and Curriculum Authority This report has been commissioned by the QCA, Contract 2686

Table of Contents

Executive Summary

This review of teacher assessment has looked at teacher assessment in practice in a number of countries to see what works best and to consider the implications for Assessing Pupils' Progress (APP). APP is an innovative approach to integrate teaching and assessment to improve and keep track of student learning. It involves professional capacity building to make teachers sensitive to the developmental progression of their students. In addition to published research evidence from other countries the review had access to evaluation reports carried out during the piloting of APP.

The emphasis of the review was to capture research evidence of the conditions under which teacher assessment works effectively and reliably. The review has shown that in assessment systems similar to the APP it is possible to gain high levels of reliability. However high levels of reliability cannot be taken for granted. Some systems have disappointingly low levels of reliability despite the implementation of training schemes for assessors.

The APP uses a well structured system with assessment focuses clearly described. The evaluation reports indicated that for most teachers the reliability of judgments based on the APP system are satisfactory for purpose. An examination of the overall distribution of levels awarded under APP compared with those resulting from external moderation and from optional tests showed a reassuring similarity. This indicates the likelihood of acceptable validity when fully implemented.

The review looks at issues that may be worth considering as the system is implemented and makes suggestions for a future evaluation strategy.

The Authors

Gordon Stanley is Director of the Oxford University Centre for Educational Assessment and Pearson Professor of Educational Assessment at the University of Oxford. He was President of the Board of Studies in New South Wales in Australia from 1998-2008. In this role he was responsible for curriculum and assessment for schools K-12, the registration and accreditation of non-government schools and for overseeing the development of standards-referenced reporting in public examinations. He is an Emeritus Professor of Psychology from the University of Melbourne and Honorary Professor of Education at the University of Sydney.

Robert MacCann is a Visiting Research Fellow at the Oxford Centre for Educational Assessment and was head of Measurement and Research Services at the Board of Studies, New South Wales for fourteen years until 2008. He directed the research programme and operational activities of this research centre in areas such as standards-based reporting, computer-based testing, school assessment and improving the psychometric properties of tests.

John Gardner is Professor of Education at Queen's University of Belfast and Visiting Professor at the Oxford University Centre for Educational Assessment. He has published widely in the area of assessment for learning and teacher assessment.

Laura Reynolds and **Imogen Wild**, from the University of Bath, are psychology work placement students at the Oxford University Centre for Educational Assessment.

Introduction

The Assessing Pupils' Progress (APP) project was initiated by QCA with the support of the National Strategies to trial a new approach to teacher assessment. The aim was to support teachers in developing their skills in assessing progress in reading, writing and mathematics within the context of a broad curriculum. The APP website provides details of the programme:
http://www.standards.dfes.gov.uk/primaryframework/assessment/app .

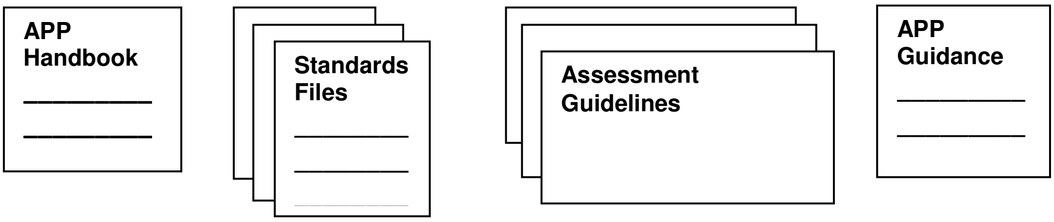
The model of assessment adopted for APP involves the periodic, systematic review of achievement as evidenced through a range of sources, including what teachers 'know' about their pupils as a result of everyday classroom interactions. Rather than being dependent on the outcomes of a 'one-off' task/test assessments are derived from a broad evidence base.

This periodic assessment of pupils' progress involves the review of evidence against criteria for each attainment target. Within each attainment target a set of assessment focuses (AFs) based on the national curriculum programmes of study have been derived to support systematic assessment of developing skills. When making assessments teachers use assessment guidelines laid out in the form of grids which illustrate performance at adjacent levels in each of the assessment focuses.

A level judgment is made for each individual assessment focus and teachers then follow a flow chart to arrive at an overall level judgment for the attainment target. Thus at designated points teachers are required to review their 'evidence' using structured assessment criteria with two outcomes:

- a profile of strengths and weaknesses across a range of assessment focuses in each attainment target to help determine next steps in teaching and learning;
- a judgment expressed as a national curriculum level for reading, writing and for each mathematics attainment target.

To assist schools and teachers in implementation a set of resources that assist the development of skills appropriate to the task of evidence-gathering and relating this to common standards is available. The set has four components:



The **APP Handbook** provides contextual information and practical guidance for APP implementation at whole-school level. The **Standards Files** contain annotated pupil work that exemplifies performance at a range of levels. Annotated student work is a common resource used by education systems to assist teachers to check their internalised standards against system wide standards. The **Assessment Guidelines** give level-related criteria for each AF and offer a simple recording format for an individual pupil. Guidelines are available in two formats. In A3 format all National Curriculum levels from 2 to 8 are on one sheet and are designed to enable teachers to spot gaps in pupils' profiles and to assist in seeing how pupils progress up through the levels. The A4 format depicts two levels on a page is recommended for recording judgements for individual pupils. The **APP Guidance Booklet** is designed to provide additional support for implementation.

The current materials available for the roll-out of APP have been developed and refined as a result of pilot studies using an action research paradigm whereby external evaluations provided feedback at each stage of development. The evaluation reports on the pilot studies provide some indication of the issues head teachers and teachers have been concerned about as the materials and process has been piloted. Clearly the team developing the process and resources has had to balance what have often been conflicting opinions and views, some of which are the inevitable 'teething' issues with any educational innovation.

APP has been trialed and is now being introduced in a context in which there has been much debate about external testing and 'teaching to the test'. Evidence presented to the House of Commons Children, Schools and Families Committee in their 2007-8 investigation on Testing and Assessment (See Testing and Assessment, Volume II, 2008) indicates that in England the testing and accountability agenda has distorted teaching practice and has lead to a narrowing of the curriculum. It is important to note that teachers involved in the pilot programme had to balance their involvement with APP while dealing with the requirements of the existing school regime of assessment and reporting. A heavy overhang of accountability will continue to provide challenges for implementation of APP, particularly if it is to be a central component of the Making Good Progress strategy.

This report will deal with the issues coming out of the development phase and then present evidence from research reports from other systems to provide some pointers for creating an appropriate basis for successful implementation of APP at a system level.

Assessment for Learning

The APP, like similar standards-based systems throughout the world, is supported by a wide range of research on the optimum ways of enhancing learning. It is underpinned by the *assessment for learning* philosophy - that assessment should be focused on diagnosing the effectiveness of learning and used in planning future learning (Black and Wiliam, 1998a; Stiggins, 2002). This philosophy has gained widespread acceptance and has heavily influenced assessment practices, particularly in the junior years of schooling.

Black and Wiliam (1998b, p.15) make some wise remarks on the improvement of learning:

> *"…the improvement of formative assessment cannot be a simple matter. There is no 'quick fix' that can be added to existing practice with promise of rapid reward. On the contrary, if the substantial rewards of which the evidence holds out promise are to be secured, this will only come about if each teacher finds his or her ways of incorporating the lessons and ideas that are set out above into her or his patterns of classroom work."*

A second feature of modern assessment has concerned the validity of such practices, giving rise to concerns about the need for authentic assessment that more adequately reflects performance required in real life situations. As Fitzpatrick and Morrison (1971, p. 268) state:

> *"…the potential value of the performance test lies in its closer approach to reality – its greater relevance in determining the degree to which the examinee can actually perform the tasks of the criterion job or some other situation."*

It has been argued that such assessments can serve as motivators of student learning, encouraging instructional strategies that foster reasoning, problem solving and communication (Frederiksen and Collins, 1989). Heavy reliance on external testing in a high stakes environment has undesirable features that may work against assessment for learning. It tends to promote "teaching for the test" (Morrison and Tang Fun Hei, 2002) and may create construct-irrelevant variance from the anxiety and low self esteem exhibited by the least successful students (Harlen and Deakin-Crick, 2003). Some students may be turned off formal learning forever. Gipps (1995) has argued that such students be assessed through multimodal tasks in non threatening settings to reduce bias in assessment.

Some of these criticisms are not unique to external testing. Teachers may teach a narrow curriculum by selecting their favorite bits for emphasis and may employ classroom assessment practices which create similar pressures (Phelps, 2008). It cannot be assumed that teacher assessment necessarily avoids the problems associated with external testing without some assurance of consistent professional standards for effective teacher assessment. Hence the importance of specified programmes like the APP.

An important feature of modern assessment practices is the concept of lifelong learning – that students should be empowered to take responsibility for their learning through self-regulation and reflection (Marzano, Pickering and McTighe, 1993; Klenowski, 1995).

By developing these patterns of operation, students can become autonomous learners when they leave the confines of formal education. They may also be encouraged to stay on longer in formal education and may possibly return to it at various stages in their lives. McDonald and Boud (2003) have shown the strong gains in examination marks that are obtainable when students receive training in self-assessment and peer assessment.

The attempts to assess more complex thinking skills have shifted the focus away from formal testing and towards gathering a wider sample of behaviours. This has led to the recognition that teacher observations could play a stronger role in assessment than hitherto thought. Moss (1994) has developed a hermeneutic assessment philosophy which argues for an integrative approach for the assessment of portfolios. She contrasts this with the “aggregative” approach from the psychometric paradigm. In the latter, marks are independently awarded for particular tasks and are then combined through a formal weighting scheme to obtain a composite. The hermeneutic approach, on the other hand, involves holistic, integrative interpretations of collected performances. This type of approach requires considerable experience on the part of the teacher and may be considered appropriate for low stakes assessment where the focus is on formative assessment.

When assessment is focused on the classroom, the professional status of teachers is enhanced. Baker, O’Neil and Linn (1993) argue that the use of performance assessment type tasks serves as a powerful professional development tool when teachers are involved in the design and marking of such tasks. In addition, teachers’ involvement in assessment moderation and standard setting are invaluable in helping them to assign performance levels correctly according to national standards.

The recognition of the importance of formative assessment practices flows from reviews that have shown considerable learning gains are possible when they are employed (Natriello, 1987; Crooks, 1988). Studies by Newmann, Bryk and Nagaoka (2001) and Boaler (2002) have shown that these new assessment approaches can improve learning. The Newmann et al study reported standardized effect sizes of 0.43, 0.52 and 0.64 for comparisons in reading, writing and mathematics, respectively. The Boaler study reported a standardized effect size of 0.21. Wiliam, Lee, Harrison and Black (2004) also reported a mean effect size of 0.32 over several measures, in studying the effectiveness of formative assessment strategies.

The APP system shares characteristics with many such systems throughout the world that integrate assessment with instruction to improve classroom learning through an assessment for learning approach. It has been developed to take account of the important research and policy papers outlined above.

Background for the APP

The Assessment of Pupils' Progress (APP) is a structured approach to the teacher assessment of pupil learning, embedded in a standards-referenced framework. As such, it is consistent with the many assessment schemes being introduced throughout the world that are attempting to report in a more meaningful manner than simply marks alone. At the heart of these schemes is the reporting of student achievement in terms of a series of verbal descriptions that indicate the characteristics of the learning – "what a pupil knows and can do". In such standard-referenced schemes, this has generally required that each supporting curriculum be specified in more detail than formerly. This has been done with the National Curriculum in England, which underlies the APP.

The APP system currently operates in three Key Stages of the National Curriculum. Key Stage 1 comprises Years 1 and 2, Key Stage 2 comprises Years 3 to 6, while Key Stage 3 comprises Years 7 to 9. These stages span a typical age range from 5 years to 14 years. At the end of each Stage, there is an *expected level of achievement* for the average student, based on the National Curriculum outcomes. The levels scale runs from Level 1 (the lowest) to Level 8 (the highest). Within each level, there are three subcategories, denoted high, secure and low. Sometimes these categories are denoted *a*, *b* and *c* respectively. Thus a student may receive a secure 3, which may also be denoted as a 3*b*.

At the end of Key Stage 1 (Year 2) the expected achievement is at *Level 2*. At the end of Key Stage 2 (Year 6), the expected result is at *Level 4*, while at the end of Key Stage 3 (Year 9), the expected result is around *Level 5* to *Level 6*. These are expected results for the typical student. Individual students may be working ahead of or behind these expected levels.

External testing at the end of these stages is in a state of flux. Since the mid 90s, a series of external tests known as *National Curriculum Tests* have been used to determine student achievement at the end of Key Stages 1, 2 and 3. In Key Stage 1, external tests in English and Mathematics are provided to schools which can administer them at a time suitable to the school. The tests are marked within the school itself, not externally marked, but samples of the marking are externally checked. From 2005, the results from the tests have been used to inform teachers' own judgments. It is these judgments, rather than the test results, that are centrally collected and which are also provided to parents.

In Key Stage 2, there is an external testing programme in English, Mathematics and Science which has operated since 1995. English comprises three tests: reading, spelling and writing. In the writing test, a short response and a longer response are required. Mathematics comprises three tests: non-calculator based, calculator-based and mental mathematics. Science comprises two tests. For Mathematics and Science, a single outcome level is provided. For English, separate results are given in reading and writing as well as a combined score in English.

These results are used as high stakes accountability measures, with the publication of *league tables* showing the proportion of students reaching Level 4 in each primary school. They are also used by the schools inspectorate to evaluate the performance of schools and are used on a national level to monitor the extent to which standard targets are reached. However, it has been proposed to replace this system with one involving shorter *single level tests* (the testing targeted at only one achievement level), which would be offered twice a year. Students could attempt these whenever they were judged to be ready, regardless of age.

In Key Stage 3, external testing in English, Mathematics and Science had been operating since 1993. However, the widely publicised marking problems in 2008 have prompted the government to discontinue this form of testing. It is expected to be replaced by a system of sampling, where a proportion of students are tested each year to measure educational standards.

This is the educational background of which the APP is a part. The APP approach fits into this wider assessment system as a school-based programme primarily designed to improve classroom assessment. It puts the teacher at the heart of the assessment process, the judgments of student progress being determined by the class teacher. A major use of these judgments is for diagnostic purposes within the school – to improve teaching and learning. The structured process gives considerable support to the teacher in how to make these judgments.

This support has an in-service function in that it provides training on relating the assessments at the school to the national standards. Apart from specific training sessions in the use of APP, the materials provided include a *Training Handbook*, *Assessment Guidelines* with a simple recording format, and *Standards Files* that give work samples of students' work which exemplify the national standards at the different levels. In addition to this support, an external moderation system was employed to further strengthen the comparability of the levels awarded across schools. Participation in these moderation meetings gave an additional in-servicing benefit to teachers.

The Pilot Programme Evaluations

Pilot project evaluation reports (see Appendix 1 for summaries of each report) provided information about teachers' experience of APP during the developmental phase of the project. Clearly those involved ranged in their views on a number of issues and attempts were made to address many issues as the project moved from pilot towards implementation. There were some common elements coming out in the evaluation reports and in teacher feedback.

In their feedback teachers reported that they found the scheme to be useful and it provided them with a better understanding of both pupils' learning and also the national curriculum levels. They valued the fact that they had more information than obtained from just using the optional tests.

While initially there were difficulties in obtaining sufficient suitable evidence, understanding of the evidence required improved, which allowed better teacher judgments and also confidence in these judgments to improve.

As a result of participation in the pilot, teachers have developed a deeper understanding of assessment and increased the range of assessment activities used. Also teachers have begun to implement assessment practice into their planning processes to enable the gathering of sufficient evidence.

Particular difficulties were found in obtaining evidence for reading assessment and also evidence for writing assessment was initially too structured with little opportunity for independent work.

There were contrasting opinions on the use of APP for whole class assessment. Some teachers were concerned that it would be far too time consuming; others believed that by using a focus group for detailed assessment they should be able to assess the whole class as they have developed a deep understanding of abilities and levels. However with whole class assessment the value of having a detailed individual assessment for each pupil was noted.

The time taken for an individual pupil's assessment ranged greatly from teacher to teacher and also within subject area. No explanation of this has been given except for a lack of understanding of what constitutes suitable evidence and also national curriculum levels. Time taken to assess each pupil was reported to fall over the course of the pilot.

There was a great range in discrepancy between the judgments given within school and those given by a moderator; a lower agreement was found for overall levels than for individual AFs. A range in time taken to prepare for moderation events was also reported and this did not decrease over the course of the pilot. Moderators reported that there was little challenge to judgments made from within the school.

No indication is given to a link between time spent on assessment and the accuracy of assessment. This needs to be studied to see if efficiency/accuracy trade-offs can be known to provide input to professional development and training for new users of APP.

The majority of parties concerned believed that APP is superior to current assessment systems. Head teachers praise the strong links that it shares to curricular planning and also to understanding pupils' learning.

External structure is critical to the success of the scheme as it encourages the completion of assessment stages by deadlines in order to continue with the next stage i.e. moderation.

Lessons from Pilot Evaluations

In examining the evaluation reports from the APP pilots a number of issues were identified which need to be part of on-going evaluation.

Manageability and Evidence Records

One of the attractive features of the strong emphasis on classroom observation and judgment in APP is that assessment is embedded in teaching practice and in the encouragement and support of student learning. Flexibility in the evidence requirements has a practical benefit in that it allows for less rigidly specified activities than occurs when standardized recording is made from commonly administered tasks. The downside is the need to record sufficient information each time a judgment is made so that it can be seen as appropriate evidence if the judgment is challenged.

Presumably uncertainty about how much detail needs to be kept as an evidence record accounted for the large range of time differences teachers gave for the workload associated with the process.

A significant issue in teacher assessment regimes is the amount of evidence collected and how it is recorded. Given that a prime focus of teacher assessment is to provide an opportunity to make judgments about student progress and to guide the next step in learning it is important that the assessment regime is manageable in terms of teacher workload and provides timely feedback (Hattie & Timperley, 2007).

Good teaching involves many verbal interactions between teachers and pupils, and between pupils, all of which provide information about the development of knowledge and skills. The dilemma is how to capture the information from these interactions without the recording process intruding and potentially distracting from what would otherwise be a simple verbal exchange. Moreover to capture an equivalent level of evidence from all class members, so that equitable assessments are made, may inevitably lead to a more structured series of interactions.

One of the dilemmas not yet resolved in the assessment for learning literature is the balance between supportive engagement with students and the observer/recorder perspective. Teacher assessment developed in an assessment for learning paradigm need not be seen as limiting pedagogical practice. Nor need it be seen as pedagogy itself. According to *Teaching and Learning Scotland* assessment as learning is about reflecting on evidence of learning (http://www.ltscotland.org.uk/assess/as/intro.asp). This is part of the cycle of assessment where pupils and staff set learning goals, share learning intentions and success criteria, and evaluate their learning through dialogue and self and peer assessment.

Clearly an important element in the success of the APP, evident in the evaluation reports, is the extent to which the school management supports APP as a priority activity for the school.

Moderation

Teachers need time to moderate each other's judgments in relation to common standards. As APP is to play a significant role in charting student progress, confidence in the consistency of judgments needs to be established. Clearly the provision of annotated work samples in the standards files provides a strong reference point for teachers to calibrate their judgments against standards. Good practice suggests that at school level some moderation exercise would also provide a check against deployment of common standards.

Validation of level judgments can be looked at in relation to external tests. However this assumes that the tests themselves are appropriate for such a task and will depend on what testing regime is in place.

Assessment capacity-building

The pilot projects demonstrate that developing teacher assessment capacity is vitally important. Clearly, evidence from these pilots shows that teachers initially differ in their subject knowledge, assessment practice and pedagogical understanding. Structured training and on-going professional support is very important. Incidentally such a finding is not surprising and mirrors the experience of Klenowski (2007) in Queensland.

Standards-based Teacher Assessment in Australia

The APP is one manifestation of a standards-based system. While there are APP features which are emphasized to suit local conditions, many other such systems are grappling with similar problems around the world. In particular, the Australian educational system, inheriting the British educational traditions, has also implemented standards-referenced reporting in each of the states.

While teacher assessment has played a significant role in Australian school assessment and reporting, to date for students it has mainly been related to a high stakes reporting environment at senior secondary level. Assessment of student performance as an accountability measure in Australia has been less prominent in the management of school systems than in England. However with the introduction of a new national testing regime, and performance contracts between the federal and state governments, this seems about to change.

The New South Wales (NSW) system

In NSW the state curriculum is organised into six Stages, with typical students in each Year expected to perform at the corresponding Stage as follows:

Early Stage 1 – Kindergarten
Stage 1 – Years 1 and 2
Stage 2 – Years 3 and 4
Stage 3 – Years 5 and 6
Stage 4 – Years 7 and 8 (high school)
Stage 5 – Years 9 and 10
Stage 6 – Years 11 and 12

In the junior years of school in NSW student progress is reported on a five-grade scale from A to E and no external moderation takes place.

The Common Grade Scale describes performance at each of five grade levels.

Grade A
The student has an extensive knowledge and understanding of the content and can readily apply this knowledge. In addition, the student has achieved a very high level of competence in the processes and skills and can apply these skills to new situations.

Grade B
The student has a thorough knowledge and understanding of the content and a high level of competence in the processes and skills. In addition, the student is able to apply this knowledge and these skills to most situations.

Grade C
The student has a sound knowledge and understanding of the main areas of content and has achieved an adequate level of competence in the processes and skills.

Grade D
The student has a basic knowledge and understanding of the content and has achieved a limited level of competence in the processes and skills.

Grade E
The student has an elementary knowledge and understanding in few areas of the content and has achieved very limited competence in some of the processes and skills.

To help teachers assign grades to their students for the purpose of reporting to parents, work samples are provided by the Board of Studies for each grade level. These are accessible to teachers and students from the Assessment Resource Centre, ARC, (http://arc.boardofstudies.nsw.edu.au/).

At this junior level, there are no grade descriptors specific to each subject – only the general descriptors given above. However, they are given meaning through the provision of work samples. For Stage 4 and under, the emphasis is on diagnostic testing to improve learning.

For a number of years there has been State wide testing which has played a mainly diagnostic role. External testing has recently been introduced in Literacy and Numeracy in Years 3, 5, 7 and 9 across all Australian state systems by the Federal Government. To date such testing has not been used as a strong school accountability measure and therefore has functioned as relatively low stakes. Nevertheless as school accountability regimes become more performance-oriented there are signs of this changing to higher stakes at least for teachers and schools.

NSW has the largest education system in Australia and has two major points of certification: at the end of Year 10 for the School Certificate (SC), and at the end of Year 12 for the Higher School Certificate (HSC). Each year over 80,000 students attempt the SC and roughly 65,000 attempt the HSC. The NSW Board of Studies is responsible for both the underlying curriculum and the conduct of these two credentials. Prior to 1998, the NSW system of assessment was norm-referenced for both reporting points, with the scaled mark distributions being standardized to fixed percentages. In 1998, the SC moved to a standards-referenced system and in 2001 the HSC followed suit. The two systems of assessment (SC and HSC) differ in how they operate, the former being low stakes for students and the latter, very high stakes.

School Certificate

External Measures

At the School Certificate, student results are a mixture of teacher assessment and external measures. Compulsory external tests are held in the subject areas of English Literacy, Mathematics, Science, History and Geography. The entire Year 10 candidature (over 80,000 students) sits for these tests. The testing covers a range of item types, including multiple-choice, short answer and extended response. Apart from the multiple-choice (which are computer-marked), all responses are written in pen and paper in answer booklets, which are marked externally in various marking centres. Standards-referenced achievement on these tests is reported in six bands, from Band 1 (the lowest) to Band 6 (the highest). An additional compulsory test, Computing Studies, is tested online with a choice of sessions over three days, the students using the computer labs at their school. As the current Computing Studies testing is multiple-choice or objective, these items are computer-marked. This external testing results in three indices of achievement:

(i) a scaled mark (/100)
(ii) an achievement band for the test (from 1 to 6)
(iii) verbal descriptions of a typical student's performance for each achievement band.

The achievement bands are obtained through a standards-setting process each year. The default method is the multi-stage *Angoff Method* (Angoff, 1971), with three stages and typically six judges. In the first stage, all item cutscores are rated independently. In the second stage, the judges confer, with the Stage 1 data available to all. Some changes in the item cutscores are usually made in Stage 2. In Stage 3, samples of students' responses around the cutscore borderlines from Stage 2 are given to the judges. A further adjustment to the cutscores may be made.

A second standard-setting method, the *Bookmark Method* (Mitzel, Lewis, Patz & Green, 2001), is being used in Computing Studies and is being progressively introduced for the multiple-choice sections of the other subjects.

The standard-setting results in five total test cutscores, for each of the six achievement bands (Band 1 not requiring a cutscore). Let C_6 denote the raw cutscore for Band 6. To ensure the scaled marks are comparable across years, each raw cutscore is scaled using the following anchor points:

$$\begin{aligned} 100 &\to 100 \\ C_6 &\to 90 \\ C_5 &\to 80 \\ C_4 &\to 70 \end{aligned}$$

$$C_3 \to 60$$
$$C_2 \to 50$$
$$0 \to 0.$$

Raw marks lying between these anchor points are scaled by linear interpolation. The level of achievement at each anchor point purports to be comparable over time.

Internal measures

The above outline describes the external measures for the School Certificate. In addition, there are school assessments that cover each course *more broadly* than the external tests.

For the School Certificate school-based reporting, each school devises an assessment policy based on Board of Studies guidelines. Generally, the school produces a student rank order in each subject, based on a combination of classroom tests, assignments, presentations and so on. To distinguish the broader school-based achievement from the external testing, a different five-grade scale is used: from Grade A (highest) to Grade E (lowest). Note that Mathematics, which tends to differentiate students more sharply than other subjects, has a 10-grade scale as follows: A10, A9, B8, B7, C6, C5, D4, D3, E2 and E1.

Each teacher then works down the rank order of their students and identifies the lowest ranked student whose work is at a borderline *A* standard. They continue this process to find the lowest ranked borderline *B*, and so on. To aid them in this process, they have a subject-specific set of descriptors, relating the kinds of knowledge and skills of students at each grade level. There is also a generic set of descriptors. In addition, work samples of students at each grade level are readily available to them from the ARC on the Board of Studies website. In this regard the process and resources provide similar support to teachers as that provided for by APP.

To moderate the accuracy of these awarded grades, one or more of the results on the external tests is used. In effect, a non-linear regression is employed. The observed school means on the selected external measure are used to predict the percentage in each grade that would be expected for a given school average achievement on the external measure. In practice, this is used to identify outliers: the top 2.5% of school groups who award too many high grades and the top 2.5% who award too few high grades, and so on for the bottom grades. This occurs before the results are finalised and released.

Once these school groups are identified, the school principal (headteacher) is informed of the results and asked to investigate within the school whether the grading pattern is justified. On many occasions they resubmit an amended pattern of grades. However, there is no compulsion to alter their grading pattern.

Sometimes they argue that the particular school subject group is outstanding in the particular subject and that this is not reflected in the school mean on the external criterion. For example, for a school subject group in French the principal may argue that their mean on the external test does not do justice to their particular achievement in French. In these cases the grade pattern is not altered, despite the fact that the school is an outlier.

Higher School Certificate

External Measures

The HSC is a high stakes system based on public examinations where the results are ultimately used for tertiary entrance selection. There is a Board of Studies reporting where the Board indices of achievement are given on the HSC record of achievement. In addition, the Universities Admission Centre (UAC) is given access to these student results and conducts a further scaling which creates a single common scale across all subjects. From this, UAC produces a general achievement ranking, the Universities Admissions Index (UAI), which is used to select which students may be admitted into particular university faculties if demand exceeds the number of places available (which frequently occurs).

First, the Board system of reporting will be considered. This system is similar to that described for the School Certificate external tests. There are six performance bands, from Band 1 to Band 6. Each band has an associated set of descriptors relating the type of performance expected of students in the band. The five cutscores associated with the six bands are determined through a multi-stage Angoff Method that operates in the same way as for the School Certificate. Further, the scaling of the cutscores to the anchor points of 90, 80, 70, 60 and 50 is performed as described for the School Certificate. The HSC reporting lists scaled marks in each course, the achievement band obtained, and the description of performance of students in each band.

Internal measures

Schools also submit assessments for each student in each course (see Board of Studies NSW, 2003). These assessments count 50% toward the final result in each course in the tertiary entrance scaling. Thus, they are high stakes measures. The Board of Studies issues assessment parameters that set down the components of the assessments and their weightings. It also constrains the number of assessment tasks that can be used, so that students are not placed under too much pressure. As a result of their internal assessment programme, schools produce an assessment mark (/100 for the usual length course), in which the rank order and gaps between the students’ marks are considered to be important information. It has been long accepted that within a school the teachers can rank their students quite accurately (for example, Elley and Livingstone, 1972). However, it is not accepted that they can accurately place their school group in relation to the rest of the state candidature. MacCann (1998) demonstrated that there was a tendency for the raw assessments to be higher than the scaled marks. Moreover, the schools that most inflated their assessments were *the lower achieving schools*.

For such a high stakes measure, public confidence requires a moderation method where the moderated assessment distribution for a school subject group closely reflects the distribution of their examination marks. The Board of Studies determined some fundamental principles that have received general acceptance. First, the rank order of the assessment is determined by the school. Second, for a school subject group, the highest moderated assessment should reflect the highest examination mark. Third, the mean of the moderated assessments should reflect the mean of the examination marks. These conditions were built into a curvilinear moderation method, using a quadratic polynomial transformation, that statistically adjusted the assessments to have a similar distribution to the examination for each school subject group (MacCann, 1996). Before the final moderation is performed, checks are made for students whose examination performance is significantly below their school assessment – they are removed from the school subject group when the moderation parameters are calculated, and inserted back into the group by interpolation. Thus, their suspect performance on the examination does not affect the moderation. Once finalised, these moderated assessments are reported on the record of achievement alongside the examination marks.

The Board of Studies forms a composite mark by averaging the scaled examination marks and the moderated school assessments. It uses this composite mark to determine each performance band (for a student in a particular course) that is reported on the record of achievement. Thus, the moderated school assessment is a 50% component of the measure determining the performance band. In addition, these composite marks are provided to UAC, where a further scaling is performed.

The UAC scaling

The following gives a brief outline of the principles of the UAC scaling. First, it linearly standardises the composite marks in every course to a common mean and standard deviation. Second, it uses an internal criterion for scaling – the general academic ability of a course candidature as evidenced by their performance over all courses attempted. For a given course (e.g. Geography), a weighted average of standardised marks *over all courses taken* is calculated for every candidate. If the mean of this weighted average is higher than the standardised Geography mean, a positive “loading” is added to raise the Geography mean. This procedure is repeated for all courses. Thus, a new set of standardised marks is generated for every course. These procedures (involving the calculation of weighted averages) are then iterated until the differences between all pairs of loadings converge. When this occurs the scaled marks are higher, on average, for course candidatures who perform well over all courses. In practice, several other refinements are implemented. For example, the general ability criterion is also used to differentiate the standard deviations of the course marks.

With the creation of a common scale, an aggregate of marks (/500) is calculated, which is converted to a percentile rank. Thus, the highest aggregate converts to a percentile rank of 100 and so on, to produce the UAI. The UAI is produced in intervals differing by 0.05 and is a major selection device for tertiary entrance in NSW.

The Queensland system

The Queensland system contrasts strongly with the more conservative approach taken by NSW. The system has its origins in the introduction of the Radford Report (Radford, 1970) which recommended the abolition of all public examinations, which had been set by the University of Queensland, and their replacement by suitably moderated school-based assessments. The Queensland Studies Authority (QSA) is responsible for the overall implementation of the assessment programme.

The QCAR Framework

This school-based system has gradually evolved over the many years of its operation. Currently, the central features of this system are expressed in the Queensland Curriculum, Assessment and Reporting (QCAR) framework. The intention of the framework is to improve the clarity of the syllabus documents and the consistency of what is taught across the state, while allowing diversity in the way it is taught. The focus is on improving the effectiveness of classroom assessment in three areas: the capacity of teachers to make informed judgments about student work against the standards embodied in the curriculum; the capacity of teachers to obtain information about student learning to inform future teaching; and the feedback given to students about their learning.
The framework comprises five components which support the teaching and learning processes:

1) A set of essential learnings in each subject
2) A standards-based approach for the essential learnings
3) An Online Assessment Bank
4) Queensland Comparable Assessment Tasks (QCATs).
5) Guidelines for the reporting of achievement.

The Essential Learnings

The essential learnings are subject-specific statements describing what the typical student should know and be able to do at various time points in their learning of the subject. The assessment of such essential learnings is organised through the identification of assessable elements.

For example in Year 9 *Science*, there are four assessable elements:

- knowledge and understanding
- investigating
- communicating
- reflecting.

The number of such elements and their nature will differ from subject to subject.

The Standards

The Standards define levels of achievement and articulate the type of performance that is required in order to awarded a given level. The levels are grades, running from Grade A down to Grade E. For each grade there is a generic descriptor which is the same across subjects. Within each subject, however, the information becomes specific where the Standards are expressed in a two dimensional grid with the assessable elements as the rows and the Grades A to E as the columns. For example in *The Arts*, there are 25 cells with the five assessable elements of knowledge and understanding, creating, presenting, responding, and reflecting crossed with the five grades A to E. Within each cell is a description of what level of performance is required to gain that particular grade for that assessable element.

Online Assessment Bank

The Assessment Bank is an online bank that supports the everyday practices of teachers by providing access to a range of quality assessment tasks. The bank is only available to Queensland teachers and tertiary education departments. It provides materials and resources across all subject areas for Years 1 to 9. Each assessment is presented as a package which includes:

- A Student Booklet – the assessment as presented to students
- A Guide to making Judgments – states what is valued in the assessment and describes the expected qualities of learning at each level
- Teacher Guidelines – gives task specific information about the *Essential Learnings* being assessed, preparation required, implementation information and feedback suggestions
- An example of an *A* level response
- Assessment-related resources where applicable – e.g. audio and/or visual stimulus required to complete the task
- Sample responses where available – student annotated responses related to the task specific descriptors in the Guide.

In addition to the Assessment Tasks, the bank provides Educational Resources for teachers. These include *professional* resources such as information and links to articles of professional interest to teachers (e.g. readings, presentations, QSA publications and professional development materials). It also provides *classroom* resources designed for teachers to use in teaching or to adapt for teaching.

The bank will also provide an online forum for informal teacher collaboration and discussion about assessments. This would contribute to furthering consistency of teacher judgments and building a shared understanding of standards.

Apart from QSA, contributions to the Assessment Bank are being made by Education Queensland (the state department of education), Queensland Catholic Education and Independent Schools of Queensland. Schools are being asked to contribute to the bank by providing student work samples.

Queensland Comparable Assessment Tasks

The QCATs are authentic, performance-based assessments that provide teachers and parents with information about student learning. They are useful in improving the consistency of teacher judgments of student achievement. The QCATs are held in early in Year 4 (assessing end of Year 3 essential learnings), early in Year 6 (assessing end of Year 5 essential learnings) and late in Year 9 (assessing end of Year 9 essential learnings).

Schools are given a *Design Brief* which details the Essential Learnings that will be tested in the QCATs. This is sent in June of the year prior to the QCAT administration. Teachers can prepare students for the QCATs by practising skills that have not been used for a significant period of time. However, it is regarded as inappropriate for teachers to rehearse the actual (or a similar) task.

The QCAT packages are delivered to schools in hard copy. Each package contains:

- The Student Task Booklet (one per student)
- The Guide to Making Judgments (to help the teacher mark it)
- The Teacher Guidelines (to support the QCAT administration)
- Sample student responses.

Each QCAT task takes about 90 minutes to complete. The teachers are helped to grade the task responses according to state standards and they are able to provide feedback to students and their parents on the strengths and weaknesses of their performance.

As they are not used for measuring teacher or school effectiveness the QCATs are seen as relatively low stakes. Their primary function is for diagnostic purposes in the classroom and to assist teachers in grading to a common statewide standard.

Guidelines for the reporting of achievement

These guidelines provide teachers with guidance on reporting student achievement and progress for Years 1 to 9. They identify the elements that need to appear on every report and the principles of school-based reporting practices.

Moderation of Senior Assessments

The discussion above has dealt with low stakes assessment. In Years 11 and 12 the Queensland students are preparing for their exit credential, the Queensland Certificate of Education (QCE). It is of interest to observe how the school-based assessments are moderated in these senior years and how they contribute to tertiary selection.

The Five Point Scale

School assessments are used in two ways in the Queensland senior system. In the first, they are reported on a relatively coarse five-point scale as follows: Very High Achievement; High Achievement; Sound Achievement; Limited Achievement; and Very Limited Achievement.

Within each school, a school moderator is appointed to oversee the entire school programme. There are also subject moderators within the school, responsible for each subject. Apart from the within-school moderation, an external consensus or social moderation takes place. This involves several steps as follows:

- Work Programme Approval – a review panel checks the school's work programme against the corresponding syllabus to ensure that the requirements have been met.
- Monitoring – the review panels consider the school's implementation of a course of study and assessment programme.
- Verification – the review panels advise schools on the standards of Year 12 achievement, based on student portfolios. Sometimes negotiation between the review panel and the school takes place in case of disagreement. State review panels also consider samples to assess statewide comparability.
- Confirmation – a further check occurs before the final certification where review panel chairs meet to examine the distributions of the levels of achievement.
- Random Sampling – the final student portfolios are randomly sampled to assess comparability after the exit levels of achievement have been awarded.

Using Moderated Assessments to calculate the Overall Position (OP)

The OP is a ranking given to a student, based on the student's average score across the best five subjects (the scores averaged being moderated school assessments). They are presented in 25 bands, from OP1 (the highest) to OP25 (the lowest). The OPs are intended for use in tertiary selection and hence are very high stakes.

Although the assessments are moderated and reported in five categories (Very High Achievement, High Achievement, Sound Achievement, Limited Achievement, Very Limited Achievement), the QSA regards these categories as too broad for calculating an OP. Instead, the fine-grained school assessments called Subject Achievement Indicators (SAIs) are used. For school subject groups of size 14 or more, these marks range from 400 (the highest performer in that subject in the school) to 200 (the lowest performer in that subject in the school). This scale refers only to the school – such marks are not comparable across schools. To gain such comparability the QSA uses an external criterion – the *Queensland Core Skills (QCS)* test.

The QCS test is designed to measure achievement on the common curriculum elements underlying the Authority subjects, independent of specific subject content. There are 49 such common elements. The QCS produces scores with a mean of 175, a maximum of about 275 and a minimum of about 75. The SAIs for each school subject group are linearly transformed to match the scores of the QCS. This is performed after ensuring that outlier students, who score atypically high or atypically low on the QCS, do not affect the scaling conversion. The QSA regards this step as calibrating the scores on the different subjects within a school onto a common scale. However, it also has the effect of undoing the results of the consensus moderation. A school subject group with an average Very High Achievement in that subject, but with relatively low scores on the QCS, would find the scaled SAIs to be much lower than expected.

After this scaling, the scaled SAIs within the school are averaged across each student's 'best five subjects'. The resulting mark is called an Overall Achievement Indicator (OAI).

A further scaling step is performed on the OAIs. For each school, the OAIs are linearly converted to match the school's distribution of marks on the QCS test, resulting in scaled OAIs. This scaling occurs for large schools (16 or more students). For small schools (fewer than 16 students), this transformation is not performed. For intermediate-sized schools (16 to 19 students), a combination of small and large school methods is used.

The scaled SAIs, OAIs and scaled OAIs are not released to the public. In the final step, the scaled OAIs are then converted to a rank that is presented in 25 bands, which gives the Overall Position (OP). The OP is an important selector for tertiary study.

The Victorian system

In Victoria, the curriculum and assessment system is administered by the Victorian Curriculum and Assessment Authority (VCAA). From the start of school (the Preparatory Year) to Year 10, there is a standards-based system based on school assessments which are related to the standards. For Years 11 and 12, the major credential for which students prepare is the Victorian Certificate of Education (VCE), which is based on a mixture of public examinations and moderated school assessments. First, consider the use of standards in the junior years.

Standards in Years 1 – 10

In Years 1 to 10, the expected achievement of students is organised into six levels. The general expectations of when students will achieve the various levels are given below:

Level 1 – Preparatory Year
Level 2 – Years 1 and 2
Level 3 – Years 3 and 4
Level 4 – Years 5 and 6
Level 5 – Years 7 and 8
Level 6 – Years 9 and 10.

The standards are called the Victorian Essential Learning Standards (VELS), which are set at a challenging level, not minimal competence, according to the VCAA. Each standard describes what students are expected to know and be able to do at that level, and *how well* they should be able to know and do it.

Strands

The curriculum is organised into three broad strands:

Physical, Personal and Social Learning
Discipline-based Learning
Interdisciplinary Learning.

Domains within Strands

Within each strand are the Learning Domains (usually called ‘subjects’ in other systems). In the first strand, the domains are Health and Physical Education, Interpersonal Development, Personal Learning, and Civic and Citizenship.

In the second strand, the domains are The Arts, English, The Humanities, Economics, Geography, History, Languages Other Than English (LOTE), Mathematics, and Science.

In the third strand, the domains are Communication, Design, Creativity and Technology, Information and Communications Technology (ICT), and Thinking Processes.

Dimensions within Domains

Each domain is then organised into a number of dimensions. For example, in English the dimensions are Reading, Writing and Speaking and Listening. In Mathematics, the dimensions are Number; Space; Measurement, Chance and Data; Structure; and Working Mathematically.

Standards are not set for all these domains *at all of the six levels*. The Standards contain learning focus statements for all domains, to assist schools in developing assessment programmes appropriate to local needs. However the use of formal standards, on which student achievement is assessed and reported, is only applied where it is developmentally appropriate. This gives the following staged introduction of standards:

Stage	Levels	Introduction of Standards in:
Prep	1	English, Mathematics; Health & Physical Education; The Arts; Interpersonal Development
Year 2	2	ICT
Years 3 – 4	3	Science; The Humanities; Thinking Processes; Design, Creativity & Technology; Personal Learning; Civics & Citizenship
Years 5 – 6	4	LOTE; History; Geography; Economics; Communication

By Level 4, Standards have been introduced in all domains.

Conveying the Standards

Progression Points

For each domain, there are detailed descriptions of the level of achievement expected at various progression points. These points are given in quarters of a level. For example, Level 2 comprises Level 2, Level 2.25, Level 2.5, Level 2.75. In English, the descriptions begin at Level 0.5, and continue through Level 0.75, Level 1, Level 1.25, Level 1.5, Level 1.75, Level 2, and so on to Level 6. At each of these points, the expected achievement in the three dimensions of Reading, Writing and Speaking is described. These detailed descriptions are quite useful to teachers in determining the level at which a pupil is working.

Assessment Maps

A second valuable aid to teachers is the Assessment Maps which are provided in all domains at the progression points described above. These are *student work samples at a given progression point* in response to a task given under standardised conditions. The *context* for this task is provided. For example, a writing prompt was read aloud to the class, followed by a class discussion. Then students were given 45 minutes to complete a writing task, working independently. The VCAA provides annotated comments on the work sample, explaining why it is an example of work at a particular progression point.

Assessment Tasks

The VCAA is in the process of developing Assessment Tasks that teachers may set for their own students. These will include a teaching and learning sequence, assessment criteria, and assessment guide. Samples of student work will be collected and will be published with annotated comments.

Gathering the Evidence

In planning activities and managing assessment, the VCAA states that the aim is for teachers to ensure that assessment is based on a variety of tasks and is inclusive of the learning needs of all students. Multiple sources of information should be used to make judgments about specific skills and depth of understanding. It is important that unexpected outcomes, both positive and negative, are also acknowledged.

The Reporting Of Achievement

Achievement is reported on a five grade scale from Grade A to Grade E. This A-E scale is linked to the Victorian Essential Learning Standards (VELS). In every school:

Grade A means a student is well above the standard expected for the year level
Grade B means that a student is above the standard expected for the year level
Grade C means that a student is at the standard expected for the year level
Grade D means a student is below the standard expected for the year level
Grade E means a student well below the standard expected for the year level.

These grades are a succinct way of reporting achievement. In addition, the report card has a section in which the teacher describes in writing what the student has achieved in relation to the standards. Three further sections describe areas for improvement/future learning, what the school will do to support the student in this future learning, and what the parents can do at home to support the student.

There are also sections for the student to make comments and the parents to make comments.

The VCE system

The VCE is a high stakes exit credential (Year 12) that is important for tertiary selection. It is based on a mix of public examinations and moderated school assessments. A ranking of student results, the Equivalent National Tertiary Entrance Rank (ENTER) rank, is derived from these marks. About 75% of the university offers are made mainly on the ENTER rank.

Generally, students in Year 12 seeking an ENTER rank will attempt three *graded assessments*. The nature of these varies somewhat from subject to subject, but for the traditional subjects, two are often school-based assessments (each /25) and the third is the public examination (/50). In Mathematics, one graded assessment is school-based (/34) and the other two are written examination papers, worth 66 marks in total.

Moderation

For the graded assessments that are school-based, the schools submit marks on a fine-grained scale. These are usually *statistically moderated* against the examination marks. This moderation sets the highest moderated mark equal to the highest examination mark scored by the school subject group. It also sets the median moderated assessment equal to the median examination mark in the school subject group, and likewise for the 75th percentile and 25th percentile. School assessments falling between these points are moderated by interpolation. Before the final moderation is performed, various checks are made to ensure the result is not affected by anomalous performances.

Calculation of the Study Score

After moderation, the marks for each set of the three graded assessments are now on a common statewide scale. Each set is then standardised by subtracting the state mean and dividing by the state standard deviation to produce scores with a mean of 0 and standard deviation of 1. Denote these scores X_1, X_2 and X_3.

A weighted sum of the standardised graded assessments is formed, with the weights being the maximum mark values divided by 100. For example, for English:

$$\text{Weighted score} = 0.25\,X_1 + 0.25\,X_2 + 0.50\,X_3.$$

These weighted scores are then converted to ranks (with the student coming first in a candidature of size *n* receiving a rank of *n*, and so on). The ranks are then *normalised* with a mean of 0 and standard deviation of 1. Finally, when these scores are linearly transformed to a mean of 30 and standard deviation of 7, the result is called the Study Score. The Study Scores are reported on the VCAA Statement of Results. As will be discussed later, the Study Scores are later scaled to produce the ENTER rank.

The Indicative Grades

The VCAA Statement of Results also reports school-based assessments from Grade A to Grade E for each of the three graded assessments. Each grade is differentiated by using '+' and '–' where necessary: the highest grade being A+. These grades are submitted by the schools and are teacher predictions of how well the students will perform on the graded assessments.

The General Achievement Test (GAT)

In addition to the public examinations, the VCAA also administers a test of general knowledge and skills in:

- written communication
- mathematics, science and technology
- humanities, the arts and social sciences.

Each of these broad areas represents a body of general knowledge and skills that students are likely to have built up throughout their school years. As the GAT is a general test, no special study is required for it and is compulsory for students sitting the unit 3 and unit 4 courses in the VCE.

The GAT is a three hour test comprising 35 multiple choice items in the humanities, arts and social sciences (1 hour), 35 multiple choice items in mathematics, science and technology (1 hour), and two extended response writing tasks (30 minutes each). Subscores are given out of 35 for each multiple choice section and out of 40 for the two essay responses.

The GAT test is a useful measure of general achievement that helps strengthen measurements in several areas. It is used in:

- improving the efficiency of statistical moderation
- identifying school results that seem anomalous in school-assessed tasks
- checking the accuracy of marking in courses that are single-marked
- calculating derived examination scores in illness/misadventure cases.

One interesting use of the GAT occurs in checking the school assessments for major works in Art, Design and Technology, Food and Technology, Media, Studio Arts, Systems Engineering, and Visual Communication and Design. The GAT is not used to override such school assessments, but to identify school subject groups that are outliers in assessment. The VCAA would then send a team of reviewers to the school to assess the work to determine whether the assessments were justified.

The Equivalent National Tertiary Entrance Rank

Tertiary entrance in Victoria is administered by the Victorian Tertiary Admissions Centre (VTAC). The first step in calculating the ENTER rank is to perform a scaling that places all the Study Scores on a common scale. This works according to the same principles as the NSW scaling. The scores in each course have already been standardised in producing the Study Scores. For a particular course candidature, the *average over all course scores* obtained by this candidature is calculated. The course scores are then adjusted up or down depending on whether the overall average is higher or lower than the particular course average. After convergence, the scores are ready to be aggregated. For each student, an aggregate score is obtained by adding:

> the best ENTER course score in any *one of the English* studies, plus
> the ENTER course scores of the *next best three* studies, plus
> 10% of the ENTER course score for a fifth study (where available), plus
> 10% of the ENTER course score for a sixth study (where available).

This produces an ENTER aggregate with a maximum possible value of 210. The final ENTER rank is obtained by converting the aggregate to a percentile rank.

Implications from the Australian systems

There are marked similarities between the Australian systems and the English system of which the APP is a part. These similarities reflect education system responses in attempting to address the rapid changes taking place in western society. Such changes include:

- an economic shift towards service-based and knowledge-intensive industries
- the creation of societies and communities characterised by social diversity, fluidity and networks where "traditional forms of authority and social identity" exert less influence
- major demographic changes and changes in the kinds of working lives that young people of today can expect, as compared to those of their parents
- advances in information and communication technologies (Bentley, 2002).

The common themes across the systems will now be discussed, along with the differences involved in implementation.

Assessment for Learning

All systems have moved to varying degrees towards a system of more authentic assessment, where there has been a shift away from sole reliance on formal tests and a broadening of the types of assessment tasks being used. Assessment is seen as assessment *for learning* (Black and Wiliam, 1998a; Stiggins, 2002), where a prime purpose is diagnostic – using assessment to identify the strengths and weaknesses in the learning of each individual, and tailoring the teaching to enhance future learning. This is a key aspect of the APP framework.

A related notion is that of authentic assessment, where the learning is seen to reflect tasks that may need to be undertaken in real life positions. These tasks are often a complex blend of various skills that require an integrated performance for successful completion. Assessment of such tasks is more readily performed through human judgment than by formal testing.

For example, the Victorian system, like the APP, advocates the use of multiple sources of information to make judgments about specific skills and depth of understanding. Some of the suggested sources include:

- negotiated tasks with negotiated assessment criteria
- self assessment and reflection
- group assessment
- portfolios
- learning journals
- observations
- presentations
- demonstrations
- peer evaluations.

As part of this widening of information, the VCAA has stated that it is important that unexpected outcomes, both positive and negative, are acknowledged.
In NSW, some of the suggested assessment strategies for Dance in Years 1–10 are listed as follows (Board of Studies NSW, 2002):

- teacher observation
- peer assessment
- skills performance tests
- performance based on a set of criteria
- research assignments, e.g. critique of performances, newspaper articles, dance journals and magazines
- worksheets, questionnaires, word puzzles, match-mates
- observation of videos, live performances, television, Bennelong Program
- dance analysis, looking at photographs, film techniques relevant to dance
- knowledge tests – history, techniques, elements of composition.

Such assessment broadening is echoed in the Queensland system.

Standards-based curriculum and assessment

All systems have sought to obtain more meaning in the reporting of student results by expressing student achievement using descriptors of what students know and can do. This has necessitated a standards-based approach which provides a commonality of setting down standards for which students can aim in their schoolwork. At the same time, schools have freedom to adapt the curriculum to local needs and freedom in their employment of teaching methods. Much work has been expended on recasting the curricula in terms of outcomes which more clearly specify student achievement.
For example, in NSW:

The standards-based Higher School Certificate offers syllabuses that set clear expectations of what students must learn and measures student performance against set standards. A student's mark in each course is reported against descriptive performance bands that show what the student knows, understands and can do. (NSW Board of Studies, 2002).

All systems emphasise the expected progression of achievement as students move through the years of schooling. In NSW, this is expressed through Stages, culminating in Stage 6, which is expected to be reached in Years 11/12. In Victoria, there are levels of achievement, culminating in Level 6, but there is quite a fine breakdown, with expected standards being provided in increments of a quarter of a Level. In Queensland, the standards expected are given for each study Year. The APP system operates in a similar fashion, with its eight levels of progression through the national curriculum.

These standards guidelines, describing what students know and can do at each stage or level, are important in supporting teachers' judgments as to the level of achievement at which a pupil is currently working.

Expanded roles for teachers

With the progressive implementation of standards-based systems throughout the world, there has been an expansion of teachers' roles in three areas. The first is that the teacher is increasingly being seen as the primary assessor in the most important aspects of assessment. The broadening of assessment is based on a view that there are aspects of learning that are important but cannot be adequately assessed by formal external tests. These aspects require human judgment to integrate the many elements of performance behaviours that are required in dealing with authentic assessment tasks. This recognition requires that the teachers be the primary assessors. It also requires that the teachers be given adequate support to make these judgments. Part of this support is provided by the attempts to clearly specify the standards of performance in the curricula. In principle, these standards should also be transparent to the students and their parents.

A second area where teachers' roles have expanded is in the setting of standards. In NSW, panels of teachers are formed to act as standard-setting judges, an important role in which they set the cutscores which define the minimal level of achievement for each of six bands. This process effectively "equates" the achievement scale from year to year, thus allowing comparisons of student performance to be made over time.

A third area where teachers' roles have greatly expanded is in the moderation of school assessments. Some Australian states make extensive use of consensus or social moderation, an operation which relies heavily on teacher judgment. For example, in Queensland, it operates at all levels of schooling, including grading the Year 12 results into the five categories. The Queensland system of collecting and evaluating portfolios for moderation is similar to the APP moderation system. Apart from Queensland, portfolio moderation takes place in South Australia, where student materials that have been marked by teachers are inspected by external moderators.

Support for Teacher Assessment

All the education systems discussed result in teacher-assessed grades (or levels) that purport to be comparable across schools. These are reported on records of achievement and may be interpreted by parents and employers who would generally make an assumption of the comparability of these grades. It is therefore important that the process of assigning students to grades or levels of achievement be strongly supported. From the review of educational systems, it can be seen that his support from the central education body can take the following forms:

1. standards guidelines which describe what a typical student knows and can do at each standards level
2. work samples at each level obtained under strictly standardised conditions
3. work samples at each level obtained under less constrained conditions
4. comparable assessment tasks that teachers can give to their own classes and mark according to guidelines
5. assessment items from a central bank with annotated work samples
6. training and involvement in moderation and/or standard setting
7. monitoring of results by external tests.

All the Australian states reviewed and the QCA provide comprehensive standards guidelines as in 1. The APP process has clear and detailed standards guidelines which compare favourably with those developed for the Australian systems. The APP system also employs Standards Files which give annotated work samples, showing typical examples of the work of students performing at a particular level.

In the list above, the provision of work samples is differentiated into two types. These are exemplified by the NSW system. In point 2, the NSW samples are produced strictly under the same conditions for all samples during a public examination – they are produced by the students themselves without any assistance from the teacher, other students or parents, under a strict time limit. These samples are linked to the performance bands (Bands 1 to 6) which are used to report the examination results. NSW also produces work samples for its ARC which are produced in schools in a more informal way. These are aligned to one of the six stages of achievement and then assigned a grade category (A to E) within each stage.

The comparable assessment tasks, devised by Queensland and Victoria, would seem to be an excellent method of teachers internalising the achievement standards, if the information is primarily kept at the school level and used in formative assessment. In theory, it should also allow an approximate assessment of the proportion of students at each school that are working at a given achievement level – if the assessment is kept low stakes. However, the very prospect of external staff from an education body looking at a school's results would be likely to gravely weaken the process. If teachers feel under threat, then there may be teaching to the task and lenient teacher marking of their own school's results.

The provision of an assessment bank, from which teachers can draw tasks to give to their pupils, seems to be an excellent way of training teachers in assessing according to the standards embodied in the curriculum. This operates in a similar way to the comparable assessment tasks, with the obvious difference that the tasks are not common across schools.

All systems provide considerable training in recognising work that is of a given standard, which is essential for teacher moderation and as standard-setting judges.

When the assessment shifts to high stakes, all systems rely on some form of external testing. This seems to be inevitable, with the external testing being seen to have an objectivity in the minds of the public. However, some systems like NSW also use it in low stakes assessment as a way of monitoring teacher allocation of grades. This works quite effectively as it is only used to identify outliers – even then, the school can argue why it wishes to retain the grades and the school has the final say.

A disjunction between high and low stakes procedures

In all Australian systems (and the English system of which the APP is a part) there is a shift in the type of procedures as one moves from low stakes to high stakes assessment. For the Australian education systems, the senior years of schooling (Years 11 and 12) are characterised by the use of *external testing* and the use of these external tests to moderate or monitor the school-based assessments. In both NSW and Victoria, the external tests are public examinations, set on standards-based syllabuses. School-based assessments are important, usually comprising 50% of the final award in each course. However, despite the fact that these assessments may be based on standards guidelines, worksamples produced under specified conditions, and annotated worksamples, the education bodies still statistically moderate them to have a similar distribution to the pattern of examination marks.

Even in Queensland, which bases its Queensland Certificate of Education on 100% school assessment, an external test is employed – the Queensland Core Skills Test (QCS). Within each school, this test converts the school assessments for the subject group to have a similar distribution to that obtained by the group on the QCS. Thus high assessments obtained by consensus moderation could be substantially lowered by a group's relatively low performance on the QCE.
If the school assessment scores in a course were accurately equated through consensus moderation across school groups, and public confidence was placed in this process, then it would be logical to merge all these moderated assessments to form a statewide distribution of assessments for the course. These large statewide distributions could then be calibrated across courses through the QCS. This, however, does not occur.

Cumming and Maxwell (2004, p. 103) point out that

“there is a disjunction between the design of outcomes-based curriculum for Preschool to Year 10 students and the curriculum of Years 11 and 12…assessment in the upper secondary school (or post-compulsory years) tends to take on a different (competitive) character because of its contribution to selection for future studies and work.”

This disjunction between the secondary junior years and the senior years is mirrored in the British system with the General Certificate of Secondary Education (GCSE) in Years 10 and 11 and the A levels in Years 12 and 13 being based on external examinations.

Teacher Assessment in Scotland and Wales

Scotland

The Scottish Survey of Achievement (SSA) is a sample survey undertaken each year (since 2005) when it replaced the National Audit and the Assessment of Achievement Programme and covers pupils in the primary classes P3 (~7 years), P5 (~9 years) and P7 (~11 years); and secondary S2 (~13 years). It records test-based and teacher assessment-based grades anonymously for a selection of students for one major curriculum area each year (e.g. English language, 2005; social studies, 2006; science and science literacy, 2007; and mathematics, 2008). It also includes assessments for reading, numeracy and other core skills though there are variations in the core skills chosen year-on-year. There is a 7-point grading scale: <A (not yet achieving A – the lowest grade), A, B, C, D, E and F. The administration each year includes questionnaire surveys of student and teacher views on a variety of assessment, teaching and learning matters.

Benefits to Teachers

With the government publishing comprehensive results from the assessment and questionnaire surveys in the year following their administration, several benefits are proposed for the SSA system (SQA 2007) including:

- Teachers being able to monitor their own and their school's assessments against the national and local authority data. They are also able to consider the views of the teachers and students who respond via the questionnaires.
- A bank of exemplar materials, the National Assessment 5-14 Bank (www.aifl-na.net) is also available for teachers to adopt or use to calibrate their judgments.
- Professional development through participation directly in the system, giving first hand exposure to the assessment process and the opportunity to liaise with a colleague to carry out, compare and discuss the grading of student work.

Appropriate Support for Students

Advice is also provided to teachers on what is considered to be appropriate (permitted) support for students during the assessments. This advice is brief and sectioned into categories of student by potential grade level (SQA, 2006) e.g. for students working at the lowest level A: "Pupils working at Level A should be given help with the choice of language, content, planning and layout" while students working at D-F should not be assisted in these areas.

Moderation

Central moderation is carried out with randomly selected scripts. The information from this is fed into the system rather than back to the schools concerned, to ensure there is no sense of accountability that might skew the process.

Comparison of Teacher Assessment and Test Grades in the SSA System

Johnson and Munro (2008) have examined the SSA data in terms of comparing teacher judgments with the test results. They specifically compared the results for reading in 2005, the numeracy (by test) and mathematics (by teacher assessment) results of 2006 and the science results for 2007 (the data are available publicly from Scottish Government, 2006, 2007 and 2008 respectively). In general they concluded that teacher assessments were higher than the test grades, with the latter showing much flatter distributions.

When test results and teacher assessments were both available for individual students, the agreement between the two classifications was 40% for reading (2005), 50% for numeracy/mathematics (2006) and as low as 22% for science (2007). The authors discuss the latter result in terms of such considerations as: the strong likelihood of a distinction in what is being measured by the two types of assessment, by the teachers having no national assessment bank materials to assist them in science assessment and by the complexity of progression in science as compared to, say, mathematics. Looked at another way, the coincidence levels of teacher assessment and test results at the expected grade levels for the four pupil cohorts: A for P3, C for P5, D for P7 and E for S2; the differences are less striking with 62%, 63%, 47% and 36% respectively. Interestingly, at P7 the E grade attracted a 47% level of agreement (same as the D grade) between the two types of assessment and at S2, the extent of agreement at F was 58%.

Wales

In Wales, recent developments in teacher assessment have stemmed from the Daugherty Report: *Learning Pathways through Statutory Assessment: Key Stages 2 and 3* (Daugherty, 2004). This report advocated the strengthening of teacher assessment (through moderation) to ensure it is sufficiently robust to address a variety of purposes including to support student learning and measure their achievement, evaluate and monitor school and system level performance. Underpinning this decision was an examination of the comparative data for the result of teacher assessments and tests (including externally prescribed tasks) for the three years 2001-2003 (Daugherty, 2009).

These comparisons revealed that teacher assessment (TA) and external test/task (TT) results at key stage 2 were relatively consistent with the TA results tending to be lower than TT but still having 99% agreement (2003) within 1 level. At local authority level full agreement ranged from 60 to 97%for 2003 but the agreement level was not consistent for individual authorities with some showing changes of over 15% in the years 2001, 2002 and 2003.

The results at key stage 3 were less consistent and showed a reversed trend with TA tending to be higher than TT for mathematics and English but lower for Welsh and Science. That said, at least 97% the TA results were within one grade of the TT results. At local authority level full agreement ranged from 52 to 99% and was again inconsistent for individual LEAs with some showing changes of over 20% in the years 2001, 2002 and 2003.

The Daugherty Report prompted the *Developing Thinking and Assessment for Learning* programme, which is being rolled out across all Welsh schools from 2009 following its piloting during 2005-8. Although national testing for seven year olds in Wales had ended in 2002, the early impact of the Daugherty review was the ending of tests for 11 year olds in 2005 and for 14 year olds in 2006. That said, however, test-based practice is arguably still a feature in some schools. In policy terms, external testing has now been replaced by moderated teacher-based judgment

Since the ending of statutory tests ‘Optional Assessment Materials’ have been circulated to all schools as an aid to consistency in ‘leveling’ the performance of students at ages 7 and 11. These OAMs are increasingly supplemented, for judging standards at the age (11) of school transfer, by portfolios of evidence gathered locally by teachers working in cluster groups of primary and secondary schools.

The Reliability, Validity and Comparability of Teachers' Judgments

In the literature of education the terms 'reliability' and 'validity' are commonly embedded in the framework of educational testing and in particular classical test theory. Reliability refers to the reproducibility of the assessment and validity to the extent to which the assessment measures what it is expected to measure.

In practice neither term has a straightforward operational meaning and there are a number of different procedures commonly used when they are reported. Both reliability and validity are context dependent measures. The operational measures used in different contexts have to be evaluated as to whether or not they are fit for the explanatory purpose intended in that context. Despite this there is considerable interest in getting an indication of what levels of reliability and validity are commonly obtained in similar situations and in finding out trade-offs between precision and efficiency of the assessment regime.

Teachers are observing and making professional judgments about the performance of their students every day as they engage in classroom activities and conversations. In principle the fact that their observations are many and occurring over a number of occasions would lead to an expectation that assessment judgments based on these observations would be more reliable than assessments made on the basis of a one-off test. In general we expect that the reliability of assessment will increase with the number of observations made.

Nevertheless concerns are often expressed about how to ensure the reliability and validity of teacher assessment, especially in an era of performance management of education systems. In recent years many education systems have relied on external testing regimes to chart student progress. External testing regimes are often claimed to be more reliable and independent even if they are seen as limiting the scope of what is taken as evidence of student achievement.

In this section we are looking at the published research on teacher assessment to provide an evidence-base to understand what makes teacher assessment regimes effective. The effectiveness of teacher assessment models needs to be evaluated in the policy context in which they are operating so that effectiveness can be measured against the purpose/s required by the education system. Much of the educational research literature on the reliability or validity of teacher assessment is embedded in contexts which may not fit well into a system wide reporting and accountability framework.

Moreover in considering classroom assessment practice one can distinguish between judgments based on formal written work, such as essays and assignments of varying structure and content, and those based on dynamic interactions in classroom performances. Inherent in different classroom teaching and learning situations are varying opportunities to observe and record information to inform judgments about student achievement. Teacher assessment regimes differ in the extent of data collection and recording ranging from detailed assessment protocols to ‘on-balance’ judgments of attainment of assessment criteria. Just as with external tests and examinations, one would expect different paradigms to manifest different degrees of reliability.

To progress our understanding of the research on teacher assessment a literature review was undertaken. The studies used in the literature review were concerned with the relationship between reliability and teacher assessment. The literature consists of articles located from internet databases. The databases utilized to find relevant literature were; Google Scholar, Educational Resources Information Centre (ERIC) databases and the British Education Index. The key words used were ‘teacher assessment’, ‘judgment’ and ‘reliability’. The abstracts were read to find those citations that included data relevant to the study, those that were not deemed appropriate were excluded and the remainder were printed. The references from the relevant articles were then inspected to locate any additional related studies.

There were several criteria that the chosen studies had to fulfill. All the studies had to include a measure of the reliability of teacher assessment. In addition they had to make use of one of two methods; the first was to compare the reliability of teacher assessment with external exams or tests and the second was to look at the inter-rater reliability between teachers.

The search resulted in forty five studies that were considered appropriate for the review. The data from each study was then entered into a spreadsheet and categorised by ‘education system’ for ease of analysis. Our finding from this review was in line with Harlen (2005) who found the literature has a varied pattern of results for the reliability of teacher assessment. In our following discussion we have selected those studies and results that seem most appropriate for considering the future development and evaluation of APP.

Reliability

In this section, the reliability of measurements is discussed. This central aspect of school-based assessments must be considered jointly with other important measurement concepts such as validity and comparability. As will be seen below, in an operational setting there may be tensions operating from simultaneously attempting to maximise each of these measurement properties. In practice, there are theoretical, practical and political constraints which inevitably require some tradeoffs. Above all, in evaluating the measurement properties of a given educational programme, the purposes of the assessment are paramount.

Before considering the measurement properties of assessments, it is important to consider an important contextual influence – the degree to which an assessment is high stakes (Popham, 1987). A useful notion is to envisage a continuum running between *high stakes* and *low stakes* assessment. High stakes assessment for students takes place in a competitive environment where the precision or reliability of the assessment is crucial. The resulting indices of achievement are typically used for selection purposes, forming a gatekeeper function for entry into sought-after tertiary courses. Consequently, the marks or grades are much scrutinized. If the reporting environment is not competitive, then a lower stakes assessment may be feasible.

With the introduction of strong accountability agendas in education systems a new stakeholder dimension needs to be considered. Routine assessments may be relatively ‘low stakes’ for the student but ‘high stakes’ for the teacher and school when the number of students achieving a given performance level is part of school performance target setting.

Any evaluation of the APP programme must consider the context in which it operates in terms of the uses made of the assessments and the degree to which they are high stakes for students and/or teachers.

Definition of reliability

In a school-based setting, reliability refers to the extent to which multiple measurements of the assessments tend to agree. It is conventionally based on the concept of a true score (or latent trait) that underlies a given measurement. In practice, each measurement inevitably contains some error that makes the observed score differ from the true score. The *classical test theory model* expresses the observed score as the sum of this true score and the error score. These errors are assumed to be uncorrelated with the true scores and other errors of measurement. If the errors of measurement are large relative to the true scores (the trait being measured), then the measurement is said to be unreliable. If the errors are small relative to the true scores, the measurement is reliable. In the psychometric model described above, the reliability is defined as the *ratio of the true score variance to the observed score variance* (Lord and Novick, 1968).

In practice, the existence of large errors of measurement may cause a student’s score to vary considerably, depending on whether a student was lucky or unlucky on the given assessment occasion. In a competitive situation, this can result in a student missing a vital cutscore and being denied entry to a sought-after position. As Lord (1977) points out in high stakes selection contexts, it is the high achieving students who are particularly affected by unreliable scores. Highly reliable measurements will make their abilities evident. Highly unreliable measurements, on the other hand, may substantially lower their observed scores. For low achieving students, the position is different. Highly unreliable measurements may inflate their scores, giving them benefits that their true achievement does not warrant.

The error component of a set of measurements may be partitioned into a number of categories. There is error associated with the transient circumstances occurring on the particular occasion – for example, fluctuations in the student's wellness or mood. There is also error involved in the sampling of the assessment task. A different task may elicit a different quality response, although the student's underlying ability has not changed.

For performance-assessment type tasks (the type mostly employed in the APP), a significant source of error is that arising from marking or rating a student's performance. Contrast this situation with that employed in multiple-choice or objectively-scored tests, where there is virtually no error resulting from the rater. Even in the best of circumstances, where expensive safeguards are used, the reliability of marking may be relatively low in relation to other forms of measurement. For example, essay tasks in large scale testing programmes may be double marked (with a third marking employed if there is a significant discrepancy) but still give much lower reliability than forms of objective testing.

How reliability is measured for test scores

If the assessments are expressed in marks, a commonly used index for calculating reliability is the Pearson product-moment correlation between two sets of scores that purport to measure the same trait. This index ranges between 0 and 1, with a value of 0.90 or higher usually deemed suitable for large candidatures. This correlation takes account of the rank orders of the two measurements and the relative gaps between pairs of student marks. It does not account for *differences in scale* between the two measures. For example, consider two assessments of the same group of children, one marked by a lenient marker and the other by a severe marker. If these agree in the rank order and relative gaps, then the correlation will be a perfect 1, even though one set of scores may be 5 marks higher on average than the other set. Thus, even though the reliability index is perfect, the assessments are not comparable. The notion of comparability will be discussed in a later section.

In the USA, the early extensive use of multiple-choice testing (with its dichotomous scoring of each test item) led to the development of the Kuder-Richardson 20 coefficient (Kuder and Richardson, 1937). This is abbreviated to *KR-20*. Rather than use different testing occasions, this coefficient estimates the agreement between sections of the test on the one testing occasion. If one splits the test into two halves and correlates the scores, a reliability index is obtained. One could repeat this process by splitting the test into all possible halves and calculating the correlation for each split. As these are correlations between *half tests*, the Spearman-Brown formula is required to "step-up" each reliability estimate for a full length test. The KR-20 is equal to the average of all these split-half correlations.

As the reliability is estimated on the single testing occasion, the KR-20 estimate is referred to as a test of internal consistency. For tests not composed of dichotomously scored items, *Cronbach's Alpha* (Cronbach, 1951) was developed. When applied to a test of dichotomously scored items, this index reduces to KR-20.

One weakness of these measures is that their size is influenced by the degree of heterogeneity of the group scores. For example if a group of high ability students (with little variance in their scores) is selected, the reliability index will be lower than that for the original unselected heterogeneous group.

The accuracy of these indices is also dependent on the underlying assumptions made in their derivation (see MacCann, 2004). Two important assumptions are those of essential tau-equivalence and errors that are uncorrelated. Novick and Lewis (1967), assuming uncorrelated errors between the parts, showed theoretically that without essential tau-equivalence, Alpha tends to underestimate the reliability, while Zimmerman, Zumbo and Lalonde (1993) showed this through computer simulation. However, if the parts have correlated errors, then Zimmerman et al. demonstrated that Alpha may give an inflated estimate. Komaroff (1997) investigated the simultaneous violation of essential tau-equivalence and uncorrelated errors, concluding that Alpha is sensitive to these opposing biases, and may either underestimate or overestimate reliability, depending on the strength of each bias. Raykov (2001, 2008) has argued that structural equation modeling techniques should be used to obtain more accurate estimates.

In addition to reporting a reliability index, it is much more informative to include an estimate of the standard deviation of the errors. This is called the *standard error of measurement*. The classical test theory model enables an easy calculation of the average standard error from the reliability index. This gives an *average* measure, constant across the mark range. However, the standard error differs across the mark range, being smallest at the extremes and largest in the middle. Feldt and Brennan (1989) give a number of procedures for calculating standard errors at different parts of the mark range. Item response theory (IRT) methods are also useful. IRT software usually includes estimates of error for both persons and items. A useful IRT index is the person separation index provided in the RUMM package (Andrich, Sheridan and Luo, 2005).

If non-numerical indices of achievement are used (e.g. levels), then reliability is often expressed as a table, showing the extent of agreement. For example, 23% of cases showed perfect agreement in the grades awarded, 51% differed by 1 level or less, 78% differed by 2 levels or less, and so on. This procedure is affected by differences in scale (comparability).

Improving reliability

The reliability of a measure may be improved in two ways. One way is to make the measuring task longer by adding similar items to the assessment. This is akin to measuring the length of a table many times and taking the average. In such a process, the errors of measurement tend to cancel. Under the parallel-tests model, the increased reliability obtained through the lengthening process can be estimated from the Spearman-Brown formula (Spearman, 1910; Brown, 1910). This is an important practical method of improving the reliability of an assessment measure. It may also be applied to the reliability of the marking process – the reliability of double marking may be estimated by using the Spearman-Brown formula with a lengthening factor of two (for the two markers). However, there are obvious practical limits to the amount with which an assessment task can be lengthened.

A second method of improving reliability is to focus on improving the measurement properties of the task itself. This may involve critiquing the tasks to remove ambiguities, or to adjust the difficulty of the tasks to make them more consistent with the average ability of the group being tested. It may also involve substituting some parts of the task with items that are inherently more reliable (e.g. short answer or multiple-choice) or refining the marking scale to obtain greater clarity of the relationship between the quality of an answer and the grade awarded.

Reliability in an APP type system

For each reporting period, the APP system involves teachers assessing the national curriculum level at which a pupil is currently working. These assessments may be based on many measures, gathered over time while the teacher is observing the students. The measures are combined, not by formally weighting different elements, but through the teacher's judgments. As a result of this process, an index is produced – the estimated national curriculum levels. To estimate the reliability of these measures requires *replication*. However, the complexities of this operation and the practical requirements of classroom teaching make it extremely difficult (if not practically impossible) to exactly obtain this replication.

One form of replication is to have teachers make assessments on two occasions and look at the degree of similarity of judgments across the two occasions. In some early work on the reliability of assessment judgments Starch (1913) had college level instructors re-mark papers written by students after intervals varying from two weeks, to several months (6 & 9), to several years (2 & 4). The papers ranged in subject content from Psychology, Mathematics, English to German. He concluded that on average there was a 4.4 mark difference on a 100 point scale between the two occasions. However the range of differences on each occasion was quite considerable and would be troubling for placement of an individual.

Some evidence about the consistency of teacher judgments with respect to UK curriculum levels was provided to us by Malcolm Hayes (Hayes, 2009). This data comes from a sample of teachers most of whom can be presumed to have not undertaken the APP training programme and hence provide some base-line about teacher judgment relative to levels.

For KS2 Maths schools are required to submit teacher assessments (TAs) for year 6 pupils to the NAA for inclusion in national results tables. Each year, the test development team asks schools to provide teacher assessments for the pupils taking the pre-test and these can be matched to the assessments submitted to the NAA.

These TAs are on the same pupils and presumably by the same teachers, provided within weeks of one another. While it may be true to say that they were provided for different purposes, they are, nevertheless, essentially replications of the same measurement and do provide some evidence of the reliability of teacher assessments.

Data from each of the 9856 matched pairs of results were available for analysis. These came from seven consecutive years but results from year to year were broadly similar and have been aggregated. Although the average levels for the two measures were very similar (4.15 for the 'live' TAs and 4.13 for the pre-test TAs) 81% were the same. In other words, asking the same teachers to assess the same pupils on two different occasions produced 19% 'misclassification'. The correlation between these two assessments was 0.81 giving an r^2 value of 0.66.

Another type of replication involves assessments made independently by two or more teachers. One could imagine a situation where a class was effectively taught by two teachers, each taking turns to address the class, both always present in the class, observing and interacting with pupils. Each teacher could then independently assign a level to each of the students and the two sets of scores could be correlated. This situation would not usually occur unless set up as an experiment.

In another scenario, a teacher could take a class for a fixed period and assign achievement levels. Then a second teacher would take over for a similar fixed period and assign achievement levels. The two sets of levels would be correlated. Apart from the practical problem of the children having to get used to the change of teachers, it could be argued that the students' relative achievements may have changed over the different time periods, which would give an under-estimation of the reliability index. Even in the two-teacher scenario outlined formerly, it could be argued that the presence of two teachers in the room (rather than the usual one), could have an effect on both the students' achievement and also the ratings of such achievement.

In practice, it is easier to estimate the reliability of marking products, such as work samples or portfolios. This type of reliability estimate has been attempted for the Queensland system where it has been shown that quite high levels of reliability can be achieved. In a study by Masters and McBryde (1994), 546 student portfolios were each double-marked independently for each of three conditions:

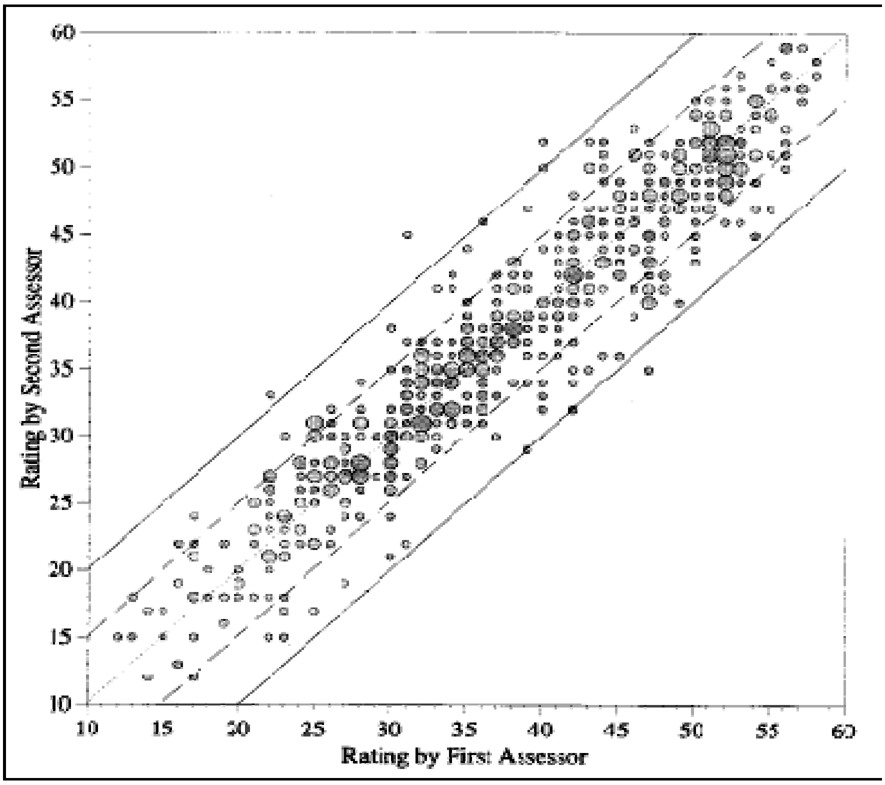
- portfolios were organised in school groups and markers assigned scores without reference to the school's assessment criteria.

- portfolios were organised in school-groups and markers were able to refer to the school's assessment criteria.

- portfolios were distributed at random to markers, who were not able to refer to the assessment criteria of the schools.

Each portfolio was marked on a 50 mark scale, first being assigned to one of the five achievement bands (Very High Achievement, High Achievement and so on) and then dividing each band into 10 divisions to create the 50 mark scale.

For each of the three conditions above, the Pearson product-moment reliability index was a very impressive 0.94. The consistency of such ratings is shown visually in the scatterplot below for the third condition of random assignment to markers (reproduced from Masters and McBryde, 1994).



The inner set of intervals on the plot shows where the differences are ± 5 marks (half an achievement level). The outer set of intervals shows where the differences are ± 10 marks (one achievement interval). From the scatterplot it can be seen that there are very few markings where the difference is more than one achievement level.

Similar results have been obtained in a Queensland random sampling of portfolios in 2006, as reported by Jordan (2008). A re-marking of such portfolios resulted in the following differences:

- 25% of folios were placed on the same rung
- 50% of folios were placed within 1 or 2 rungs
- 20% of folios were placed within 3 to 7 rungs
- <5% were placed on more than 8 rungs

This reinforces the findings of Masters and McBryde, that the marking of portfolios can be highly consistent.

A Victorian study reported by McCurry and MacKenzie (2006) also reported high levels of consistency in teacher judgments. In this study, teachers assessed students on nine generic skills, using global impression judgments made in normal teaching programmes, rather than using special generic assessment tasks. The generic skills included Written Communication, Oral Communication, Working Mathematically, Gathering Information, Problem Solving, Planning, Using Teamwork, Understanding Technology and Cultural Understanding. An eight-point scale was used. The results for pairs of teacher ratings may be summarised in the following table:

Generic Skill	% no difference	% at ± 1 point	% at ± 2 points
GS1	31.4	64.4	90.0
GS2	28.6	62.6	90.0
GS3	31.8	65.0	89.7
GS4	29.6	62.8	90.3
GS5	28.7	62.8	89.3
GS6	27.4	59.9	85.7
GS7	29.9	60.0	90.0
GS8	28.2	62.5	88.6
GS9	27.2	60.8	88.4
Mean:		62.3%	89.1%

These results compare favourably with those obtained in the double marking of essay questions in public examinations.

In the US the Early Literacy Profile, created by the New York State Education Department, provides an interesting comparison with the APP. It aims to provide a classroom-based performance assessment that is useful for accountability (Falk, Ort and Moirs; 2007). The profile developers built the assessment on daily classroom practices that supported and informed the teachers' instruction. Three profiles are assessed – Reading, Writing and Listening/Speaking. The challenge was to create a sufficiently small number of assessment tasks that made it manageable to assess in the classroom, but at the same time to provide a reliable and valid assessment.

The Reading evidence has four parts – a reading interview, reading diagnostic tools, a list of texts to be read independently by the student, and a written reading response to a text. After collecting data on the performance of each student on each of the four tasks, the teacher makes a holistic judgment on the student's level on the Reading standards scale. Similarly, the Writing and Listening/Speaking profiles are each assessed with a small number of standardized tasks that are embedded in daily classroom practice.

Falk et al. (2007) analysed the reliability of assessment for Reading and Writing using generalizability theory. For single scorer reliability, they obtained estimates of 0.74 (Fall) and 0.68 (Spring) for Reading, while for Writing, the estimates were 0.68 (Fall) and 0.73 (Spring).

In Sweden, Lindstrom (2007) investigated the reliability of marking portfolios in Visual Arts. Criteria were developed showing the particular characteristics displayed in moving from a Novice to an Expert in four levels, with each level differentiated by a “plus”, “middle” and “minus” judgment. This gave a 12-point scale. A total of 458 portfolios were marked by the classroom teachers and then independently re-marked by external teachers who taught students at the same age at different schools. The portfolios comprised a final product, sketches and drafts, reflections in logbooks, models used as sources for inspiration, and a 10-15 minute video interview with each student. On the 12-point scale, a high degree of accuracy was obtained – 78% of markings differed by two steps, and 90% of markings differed by three steps.

Such results show the levels of reliability that are possible for assessment systems where teacher training in assessment has been strongly embedded. However, some systems have reported poor to moderate reliability. In the Vermont Assessment program, the reliability of marking student portfolios in Writing and Mathematics was assessed (Koretz, Stecher, Klein and McCaffrey, 1994). The results gave reliability estimates for particular dimension scores in Writing of between 0.39 and 0.52, while the total score reliabilities in Writing were ranged from 0.49 to 0.63. For the Mathematics portfolios, the dimension reliabilities ranged from 0.42 to 0.65, while the total score reliabilities ranged from 0.53 to 0.79. These are disappointing results, which would compromise the validity of the assessments. In a study of the assessment of students’ research skills (Stokking, van der Schaaf, Jaspers and Erkens; 2004), low rater reliabilities for teacher assessments were also found.

In the US State of Nebraska the School-based Teacher-led Assessment and Reporting System (STARS) school districts identify how they will measure and report student performance on content standards. The system allows for considerable flexibility in how assessments are carried out. Brookhart (2005) had 30 maths and reading assessment portfolios double scored by trained assessors and found exact agreement on category (low/medium/high) was 73% for maths and 60% for reading. She recommended professional development for teachers on aspects of reliability including sufficiency of information and scoring procedures.

As a rule, the reliability of teachers’ judgments in the classroom setting has been generally accepted. In a review of the teacher assessment literature, Harlen (2005), found that the reliability of teacher assessments has been found to be consistent with the reliability of traditional tests.

This acceptance stems from the fact that teachers are continually observing their students over a year of teaching and are able to collect considerable information both formally and informally. The collection of this mass of information tends to increase the reliability of assessment using the same principle that operates for the Spearman-Brown formula. In the early 70s, Elley and Livingstone (1972) pointed out that there has been general acceptance of the accuracy of teacher judgments in terms of their rank order, an acceptance that has led to school-based assessments being an important part of even high stakes assessments for tertiary entrance. The reliability of the APP assessments as reported in the evaluation studies appears to be of the same order as those of the similar systems operating in Australia and elsewhere. As in these other systems, this should be continued to be verified by external sampling studies.

Validity

A second dimension on which to evaluate a set of measurements is their validity. In the past (APA, 1966), this has been conceptualised in a number of ways, which are briefly described below.

Content Validity

This is assessed by how well the material appearing in an assessment task samples the content in the syllabus or population of material to be learned. In practical classroom situations it is often addressed by constructing a table of important knowledge and skills from the population of such elements that are desired to be measured by the particular assessment task. For example, this table may have knowledge components as the rows and skill components as the columns and may indicate the weightings given to each cell. The validity of the assessment is then addressed by inspecting that the assessment instrument does in fact reflect each of the components in the table.

With the close link between AFs and curriculum objectives from the National Curriculum, the APP has been designed to manifest content validity.

Criterion-related Validity

This is evaluated by comparing the scores on the assessment task with those obtained on an external criterion, the latter being considered to provide a more direct measure of the characteristic under consideration.
Comparing level classifications from APP teacher assessments with those obtained from optional or key stage tests would produce a criterion-related validity measure.

Predictive Validity

Predictive validity is evaluated when future performance on a criterion is predicted from a current assessment task.

Concurrent Validity

This form of validity indicates the extent to which the assessment task estimates a student's present standing on some criterion.

Construct Validity

This is evaluated by investigating what qualities a test measures by determining the degree to which constructs or explanatory concepts account for performance on the test.

The criterion-related, predictive and concurrent validities have usually been evaluated by correlation coefficients or regression coefficients. Validity is sometimes assessed by the Campbell and Fiske (1959) multitrait-multimethod matrix. As Messick (1989) points out, construct validity subsumes the other categories listed above it, as content relevance and criterion relatedness contribute to the qualities that underlie a set of scores. Factor analytical or structural equation modeling techniques are often used to evaluate construct validity.

Messick (1989, p.13) has broadened the concept of validity to make it dependent on the inferences made from a set of scores:

"Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment."

Thus on this definition, there can be no general declaration that the scores resulting from an assessment task are valid – someone in the future may use them to make an invalid inference. That is, the validity is specific to a particular inference.

While Messick's framing of validity is difficult to fault on logical grounds, it does not easily lead to practical ways to evaluate validity. In practice, systems using portfolios have usually resorted to estimating criterion based or predictive validity. These have been somewhat disappointing. Koretz, Stecher, Klein and McCaffrey (1994) found only moderate correlations (between 0.47 and 0.58) between writing portfolio scores and other writing assessments in Vermont. Gearhart, Herman, Baker and Whittaker (1993) found almost a zero correlation between portfolio scores in writing and the scores on a standardized writing assessment. Shapley and Pinto (1995) obtained low correlations (from 0.17 to 0.39) between reading/language arts portfolio scores and standardized reading and language subtest scores. Shapley and Bush (1999) examined the relationship between students' portfolio ratings and their relevant scores on the Iowa Tests of Basic Skills (ITBS) in reading, language and mathematics. The correlations of the portfolio ratings and the test scores were again disappointingly low.

These low relationships may have resulted because the portfolios are measuring valuable information that is not being captured by the test scores. Alternatively, it is possible that marking of portfolios may be captured by less relevant information such as neatness of presentation etc.

The current unified theory of test validity developed by Messick (1989) has construct validity at its centre. Lissitz & Samuelson (2007) have proposed a taxonomy which is based on four elements. They consider whether the focus of the assessment is internal or focused on constructs that are external, and whether the emphasis is theoretical or practical. They argue that content validity or internal validity should be acknowledged as the critical initial characteristic to consider when evaluating the quality of an educational assessment (p.446).

The usefulness of their taxonomy can be illustrated in looking at how teachers can establish whether their classroom assessments are valid or not. The teacher should look at the planned assessment task in terms of the specification or assessment focus/ses that governs the design of the task. Then when the assessment has occurred the teacher might look at the relationships between component tasks to see whether there is consistency and whether or not the outcomes reflect the curriculum objectives.

In a similar vein the APP assessments address the issue of validity directly. The curriculum guidelines specify that a typical student at a given performance level has certain knowledge and skills and can do certain things. An observer of a group of students who have been assigned a particular performance level by the APP process may quite reasonably infer that these students can actually perform at a level implied by the level descriptors. This inference would be valid if this is generally the case. Thus the degree of congruence between the actual performance of students and their level descriptors is one test of validity.

There is a broader set of inferences that people may wish to draw. The APP assessments are part of the authentic assessment movement that attempts to broaden the basis of assessment to reflect tasks that are of real world importance; e.g. oral skills, teamwork, planning skills. To the extent that APP assessments actually measure these areas considered important in the real world, the assessments are valid for this inference.

The validity dimension is regarded as fundamental to the measurement process (Messick, 1995). A high reliability is not sufficient for good measurement if the validity is low. For example, one could envisage administering a highly reliable multiple-choice test purporting to measure dance performance and finding that the results bear no relation to other external ratings of the performances themselves.

It is in this important area that the APP process is designed to improve measurement by focussing on a wide range of classroom behaviours that are generally considered important in later life. An independent check on what teachers actually include in their assessments would enable the validity of this inference to be evaluated. The APP process also provides measures that describe what students know and can do. Independent ratings of what students know and can do would allow the validity of inference from the performance levels to be evaluated.

Comparability

The comparability of measurements indicates the degree to which the marks, grades or other indices of achievement are on a common scale. When grades are comparable, one can compare an *A* obtained in one school with an *A* obtained in another school and regard them as roughly equally meritorious. The process of adjusting school assessments (whether marks or grades) so that they are on a common scale across a population of schools is referred to as moderation. Two forms of comparability are often distinguished, one very strict and the other more approximate, with conditions relaxed. In the strict form, two different sets of measures (say *X* and *Y*) are said to be *equated* (Braun and Holland, 1982; von Davier, Holland and Thayer, 2004). In order for *X* and *Y* to be equated, three conditions must hold:

1. The two measures must be measuring the same trait (for example, achievement in Mathematics). Equating cannot take place where the traits are substantially different (for example equating Mathematics scores to Science scores).

2. The measurements of *X* and *Y* must be equally reliable. Although the statistical mechanics of an equating process may be performed that supposedly equate the two measures, the resulting scores are not strictly equated unless they are equally reliable.

3. The equating conversion relationship holds for all subgroups in the population of interest. That is, there is not one equating line for one subgroup and a different equating line for another subgroup. For example, if one tried to "equate" English scores to Mathematics scores in two separate operations – one for males and one for females – then the equating relationships for the two groups would almost certainly differ.

One can see how the above conditions for equating incorporate the key notions of validity (Condition 1) and reliability (Condition 2). The breakdown of Condition 3 is also often related to the failure of the validity condition. Nearly all cases of moderation cannot satisfy the above equating ideal. In such cases, operations that convert measures on one scale to that of another are said to calibrate the two measures (Angoff, 1971). However, renaming the process for failing to meet the conditions does not solve the essential problem. The more a moderation process fails to satisfy the strict equating conditions, the more ambiguity exists in interpretation when comparing two "equated" or "calibrated" scores.

The accuracy required of a given moderation process bears a practical relationship to the degree to which the measures are high stakes. In high stakes situations, the accuracy of the scores under comparison is vital. It is important to closely satisfy both reliability and validity considerations, as shown above. Apart from the accuracy of the scores, there is also the matter of public confidence. Many high stakes programmes for school leavers (which determine tertiary entrance) are based on both external measures to a school, such as public examinations, combined with moderated school assessments. The presence of external measures reassures the public and at the same time, the combining of two measures (external plus moderated internal) should improve both the reliability and validity of the scores or grades. In addition to measures that have high reliability and validity in a high stakes context, it is necessary that the moderation outcomes be accurate in order to capture public confidence. Moderation models may be grouped into two broad categories – statistical moderation and social (consensus) moderation. These models are placed below into a more general set of linking models.

Linking Models

In an important paper, Linn (1993) ranked five ways of linking two sets of scores, which include both statistical and social moderation. These are given below.

1. *Equating*. The two sets of scores must measure the same construct, be equally reliable, and produce a unique equating line.

2. *Calibration*. The two sets of scores must measure the same construct but may not be equally reliable.

3. *Statistical moderation*. The two sets of scores may not exactly measure the same construct, but are placed on the 'same scale' by adjusting their distributions to be similar.

4. *Prediction*. The two sets of scores do not measure the same construct but one set can be predicted from the other, the predicted scores having a reduced variance due to the imperfect correlation.

5. *Social moderation*. The two sets of scores do not measure the same construct but are linked judgmentally.

Linn presented these methods as “listed in order of statistical rigor, with equating being the most rigorous and social moderation the least rigorous” (p.85).

Statistical Moderation

In statistical moderation, a criterion is used that is closely related to the assessments to be moderated. In practice, the criterion is usually an external test or public examination that is attempted by the students in addition to being assessed within the school. If the distribution of criterion marks obtained by the school group has a similar distribution shape to that of the school assessments, it is appropriate to use a linear conversion line. Linear models of statistical moderation have been discussed in Greenall (1949), Howard (1958), Pilliner (1958) and Linn (1966). The most commonly used model is one that converts the raw school assessments to have the same mean and standard deviation as the scores obtained by the school group on the criterion. Thus, the adjusted assessments will have a similar distribution to the criterion scores. This linear conversion preserves the rank order of the school assessments and the relative gaps between pairs of students.

This may be expressed mathematically in sample notation as follows:

$$Y = \overline{Y} + \frac{S_Y}{S_X}(X - \overline{X}) \ ,$$

where X is the school assessment,
Y is the moderated assessment,
$\overline{X}$ is the raw assessment mean,
$\overline{Y}$ is the criterion mean,
S_X is the standard deviation of the raw assessments,
S_Y is the standard deviation of the criterion scores.

In competitive situations the linear model often fails to work satisfactorily as teachers may negatively skew the school assessments (with an upper score “hump” and a lower score “tail”) by setting assessment tasks at which the students can excel. However, on the more searching external criterion the school’s mark distribution may be symmetrically distributed or even positively skewed. Applying a linear model in this case results in lower moderated assessments at the top end of the scale than expected. The top assessed students may find that their moderated assessments fall well short of their criterion marks.

Conversely, if the school assessments are positively skewed (with a long “tail” for the upper assessments) but the criterion scores are symmetrical or negatively skewed, then the top moderated assessments may considerably exceed the top criterion marks scored by the school group. Very able students in this latter school may be at an advantage compared to very able students in the former school.

An alternative is to protect the moderated assessments of the top assessed students by establishing as a principle that the top assessed moderated assessment at the school shall be set equal to the top criterion mark scored in the school group. This principle usually necessitates that a curvilinear line of moderation be adopted. One useful model is the polynomial equation of degree 2, which can fix this top moderated assessment and, at the same time, set the mean of the moderated assessments equal to the mean of the criterion for the school group scores. This may be written:

$$Y = aX^2 + bX + C$$

where *X* is the raw assessment, *Y* is the moderated assessment, and *a*, *b* and *c* are constants for the school group, to be estimated from the school group data. (See MacCann, 1996 for details of the estimation). The curvilinear moderation retains the rank order of the teacher’s assessments, but does not preserve the relative gaps between pairs of students over the whole mark range. Within a restricted mark range, the relative gaps between pairs of students are only approximately maintained.

In any statistical moderation there are usually statistical tests for students with aberrant performances that might distort the moderation. These students are removed from the moderation when the school group parameters are calculated and inserted back into the group when the moderation is performed.

Comments on statistical moderation

A major benefit of statistical moderation is its transparency and reproducibility. Usually the criterion is the set of school group marks on a public examination. These are known to the school staff. Secondly, the raw assessments are provided by the school staff and are thus known to them. Thirdly, education systems usually publish the statistical algorithms for moderation, so that the school mathematics department can actually reproduce the calculations. In the NSW system, before the results for tertiary selection are finalised, teachers have the opportunity to carefully study the moderated assessments in relation to the examination marks. If they identify what appears to be an anomaly, they can refer this as an Anomalous Enquiry to the educational measurement section at the Board of Studies. Each case is considered on its merits and some cases result in adjustments to the moderation.

In statistical moderation, there is a tension concerning the degree of the relationship between the criterion and the assessments. For the moderation to be considered valid, it is desired that there be a strong correlation between the assessments and the criterion. For example, one could hardly justify using measures of athletic ability to moderate assessments in the subject English. However if the correlation is too high, then the criticism can be made that the assessments may not be measuring the unique qualities that can be captured in a school-based assessment, but merely mimic the written examination.

Related to the previous point, in a high stakes competitive environment the correlation has an important effect on the formation of any composite mark. Suppose a composite mark was formed by averaging the examination marks and the moderated assessments. Then in a school with a high correlation, the variance of the composite will be larger than a school with a low correlation. This occurs because students who are near the top on one measure will also tend to be near the top on the other measure if the correlation is high. If the correlation is low, this effect will be greatly reduced. Therefore in a high stakes situation, it is of benefit to the most able students if the correlation is high as both scores will then tend to be high. This effect may provide additional pressure for the assessments to be narrowed so that they are similar to the written examinations.

Social (Consensus) Moderation

Social moderation describes a family of procedures for aligning school assessments through a process involving professional judgment. The school assessments are associated with work samples and it is the latter that are compared across different schools by referring them to a standards-referenced scale. Frequently, the work samples are portfolios of student work that are assigned to a certain performance level by the school and are then referred to external moderators to determine whether the assignments are accurate or too lenient or too harsh. These external moderators are sometimes teachers from another school or panel members on state or district moderator panels. If the latter, they are usually acknowledged as experts in judging according to the standards scale.

The social moderation process can become quite complex and resource intensive, particularly in a large geographical area. The Queensland social moderation procedures are highly regarded and have been fine-tuned over many years. At the senior level, this moderation involves a number of steps which are set out below (Queensland Studies Authority, 2005).

Work Programme Approval

The QSA describes this step as a process of moderation. Here the relevant review panel checks the school's work programme against the corresponding syllabus to ensure that the requirements of the syllabus have been met.

Monitoring

In the monitoring process, the review panels consider the school's implementation of a course of study and assessment programme after approximately half the course has been completed.

Verification

Verification is the process by which the review panels advise schools on the standards of Year 12 achievement, based on student portfolios, in relation to the syllabus descriptors of standards. That is, it answers the question "How appropriate are the school's judgments about the achievements of its students?"

If the review panel cannot substantiate the school's decisions, then a process of consultation and negotiation between the review panel and the school take place.

Following verification, state review panels meet to examine sample verification submissions from schools in each school district. The aim is to ensure the comparability of the moderated achievement grades across the state.

Confirmation

Between the time that the school's proposed grades are received and the time of printing the final certificates, a further check is performed. The review panel chairs of the district and state panels, and standards and assessment officers from QSA meet to examine the distributions of the levels of achievement. This may involve further review of student folios.

Random Sampling

In this process, student exit portfolios are randomly sampled to assess comparability after the exit levels of achievement have been awarded. The student work is reviewed by panelists in "non-home" districts.

These processes are followed for the senior level assessments in Years 11 and 12 for the purpose of producing the performance bands.

The research literature often assumes that a middle to high correlation between two sets of ratings for a sample of portfolios is sufficient to establish that the portfolios have been correctly aligned. This is not the case, as pointed out by Linn et al. (1992). Supovitz, MacGowan and Slattery (1997) studied the mean scores allocated to portfolios in Reading and Writing. They found that for all six areas studied, the classroom teachers awarded a higher mean than the external markers, although the small sample sizes prevented some of these from being statistically significant. A very similar result was found by Shapley and Bush (1999) across ten areas for children in kindergarten and Grades 1 and 2. In all ten areas, the classroom teachers awarded higher scores than the external markers. The effect sizes reported were quite large, with eight of the ten being greater than 0.4. The authors report that “the discrepancies…may reflect teacher bias; that is, the classroom teachers are more lenient in assigning scores” (p. 119).

These results may reflect a natural tendency for the classroom teacher to see their students in the best possible light, despite the training they received in aligning the portfolios to the statewide standard scales. It emphasizes the need for external moderation of some sort, the rigour depending on the degree to which the assessments are high stakes.

Comments on Social Moderation

The type of moderation programme that is implemented for a system depends crucially on the purposes of the assessment. As has been shown in the review, for high stakes competitive assessments there is a dominant tendency to rely on external testing to moderate or monitor the assessments. The external measure is often a public examination that has the confidence of the public and is seen to be an objective assessment with external marking of student scripts, the identity of the students not being known to the markers. The weaknesses of such a system are well known – a narrowing of the curriculum where schools “teach to the test” and important educational outcomes are not measured. However, these features are predominantly determined by the high stakes nature of the assessment. If the assessment is high stakes, then no matter what assessment or moderation system is devised, *the participants will use every device available to maximise their outcomes,* regardless of whether such strategies are considered good educational practice.

In the junior years of schooling, there is an opportunity for assessment to be conducted in a low stakes context. If the assessment is low stakes, then the level of comparability needed under a social moderation need not be as high as under statistical moderation. In the latter, the assessments are usually presented in a fine-grained scale and the assessment distribution is closely matched to the criterion distribution. However in the junior years of all standards-referenced systems reviewed, and in the APP system, the assessment process groups student performance into broad categories. For example in the APP for say Year 6 students, most of the children will be working at Level 4 – whether a High 4, Secure 4 or Low 4. This sorting of *most students* into only three categories should not present major difficulties for a social moderation system. There will naturally be some misclassifications, but these can be corrected as further evidence comes in as the child passes through schooling.

The APP Evaluation Report 10 (See Appendix 1) presents data showing the distribution of levels assigned to a sample of students using the APP process compared to the distribution obtained through an optional test. This comparison is shown below for the three areas of reading, writing and mathematics.

Reading: Distribution of levels assigned by APP and optional test

Level	APP reading	Optional reading test
2	6.8% (38)	6.6% (38)
3	29.0% (166)	20.1% (115)
4	47.5% (272)	43.5% (249)
5	16.8% (96)	29.8% (171)

Writing: Distribution of levels assigned by APP and optional test

Level	APP writing	Optional writing test
2	10.8% (62)	11.0% (63)
3	37.7% (216)	38.6% (221)
4	38.9% (223)	42.1% (241)
5	12.6% (72)	8.4% (48)

Mathematics: Distribution of levels assigned by APP and optional test

Level	APP mathematics	Optional maths test
2	7.4% (43)	13.0% (95)
3	38.3% (223)	36.2% (211)
4	40.3% (235)	36.2% (211)
5	14.1% (82)	14.6% (85)

These comparisons show promising results for the APP process. For the total sample available for the study, the APP method has given similar overall distributions to that of the optional tests. Further work is needed to look at the results in individual schools to see if the school awarded levels are correctly aligned relative to each other.

In this assessment context, the purposes of the assessment should be emphasised. The assessment is *for* learning. The major purpose is diagnostic – to assess what is well-learned, and to determine how to structure future teaching to enhance learning. As noted previously, this involves the teacher at the heart of the process, gathering a wide variety of information from both formal and informal measures and forming an integrative judgment of the current level at which a pupil is working. This process empowers the teacher and enhances the sense of professionalism of the teacher. Seen in this context, the social moderation process contributes to the teacher's professional development and fosters a sense of collegiality, all of which directly impacts on classroom assessment.

Practical issues in managing the assessment process

In any assessment process that is classroom focused there are inevitable tensions which arise when implementing the assessments. There are three important factors that interact:

- The range and quantity of work on which teachers' judgments are made
- The manageability of making such judgments during teaching
- The recording and storage of evidence.

To maximize the validity and authenticity of the assessment, there is an expectation that a teacher's judgment should be based on observing a student's performance on a wide range of activities. In this way, a student is given every opportunity to show their level of functioning in relation to the national standards. Given this wide range of observations, the tension arises as to the manageability of recording it.

One possibility is for the teacher to take notes on every observation that might contribute to an assessment. This has the virtue of giving a complete picture of the student over the full range of educational activities. However in a large class, the teacher may become overwhelmed by the sheer amount of data being collected. In addition, the students may feel that they are always under observation. These effects may interfere with the natural teaching process.

Another possibility is to make observations as an everyday part of the teaching process, not writing them up as they occur, but letting the cumulative effect of such observations inform a judgment of the level of functioning of the student. This more informal approach is not dissimilar to how teachers would normally operate. They could then make some notes, when an opportunity occurred, as to some memorable observations that they felt should be recorded. Under this scheme, the note-taking would be far less extensive.

A continuum of observation/note-taking exists between the two examples listed above. An important factor which influences the degree of note-taking is whether the assessment is high or low stakes. In a high stakes environment, the level may be challenged by the student and/or parents. This would require fairly extensive storage of evidence that could be used to justify the level assigned. This could ultimately prove to be unmanageable and may prove a burden to teachers.

In a practical classroom setting, Hay and MacDonald (2008) show how teachers may adapt to dealing with the mass of information that is potentially available. The setting is the Queensland senior school curriculum in Physical Education, where the assessments are based on a portfolio system in which evidence of student learning is collected continuously across the two years of the course. This evidence in the portfolios is selectively updated, acknowledging that evidence collected at an earlier stage might no longer be representative of student achievement. The teachers in the study claimed that the statewide criteria and standards had become sufficiently internalized for them to make judgments of the student achievement without reference to the former. Hay and MacDonald describe the process as follows:

"…the teachers relied on memory as evidence upon which assessment decisions were made; there was a sense of teachers possessing their own implicit standards; perceptions of the students' attitudes and behaviours were incorporated in the teachers' assessment of students' performances; and the teachers indicated that their knowledge of the student strengthened their ability to assess them." (p.162).

This process may be an inevitable adaptation to having to assess a mass of data on each individual. The dangers are that the teachers may be influenced by extraneous factors, such as affective characteristics, in their interactions with each student. Hay and MacDonald remark:

"It is argued …that such construct-irrelevance compromised the construct validity and possible inter-rater reliability of the decisions made and advantaged some students and marginalized others on the basis of characteristics that were not specifically related to the learning expected from following the syllabus." (p.153).

These unintended effects suggest that perhaps some of the tasks set for assessment could be standardized to assist teachers in internalising the standards and to ensure that pupils, both within the same school and across different schools, are being assessed on the same attributes for at least some key tasks.

One way to preserve the best features of complete teacher observation and manageability of the evidence would be to record and store scores on only a limited number of performance tasks (perhaps three or four per attribute being assessed). These scores could be stored but the final judgment of the level at which a pupil is functioning would not be made by a formal weighted composite of the scores. Instead, the teacher would draw on all the informal observations made during teaching and integrate them with the scores on the nominated tasks. A final holistic judgment would be made by the teacher. Thus, if the student had under-performed on some of the tasks, yet the teacher had observed the child to be adequately functioning on the attributes during normal classroom work, the child could be assigned to the appropriate level. That is, the teacher could override the scores on the tasks. If called upon to justify this decision, the teacher could then refer to the observations that justified the override. Such a procedure would retain the sense of empowerment that teachers experience through working with the APP process.

In similar systems to the APP, the workload involved can become a significant issue. Falk, Ort and Moirs (2007) report that about 20% of field-test teachers reported that the process was difficult to perform in a reasonable amount of time. This difficulty mainly occurs where students have to be assessed individually, rather than as a group – for example, as in reading assessments. It is important that the APP strike the right balance in recording and storing information so that the natural teaching process is enriched and informed by the information.

Task standardization v. Freedom in teaching

A further factor that interacts with teacher workload and the assessment for learning approach is the degree of standardization of any assessment tasks. Koretz et al. (1994), in discussing the implementation of the Vermont assessment system, have argued that there is a tension between the two goals of quality measurement and improvement in teaching. They remark:

> "*Although often ignored in the policy debate, the tension between these two goals is one of the most important issues confronting the performance-assessment movement, and it will become more critical as programmes move toward greater reliance on unstandardized, instructionally embedded tasks."* (p.13).

Koretz et al. point out the large variations in teachers' implementation of the programme, particularly in the selection of tasks and the rules for revision of the products. For example, if two students produce similar quality work in their portfolios, but one of the students has been given greater opportunities for revision and more structured help from the teacher in revising the product, then the similarity in the final products is misleading.

Similarly, if teachers assign variants of similar looking tasks that differ substantially in difficulty, then students assigned the more difficult task may be at a disadvantage. On a simpler level, this problem is sometimes encountered in formal public examinations where optional questions are allowed. Sometimes one question turns out to be more difficult than the other optional questions, and the marks are lower, despite other evidence (e.g. performance on the total paper) that the students attempting the difficult question are well above average. If this effect operates across different schools in the tasks comprising work samples, then products appearing to be of similar quality may mask substantial differences in achievement.

However, the tension that occurs is that good teaching *may require variations* in both the degree of difficulty of the task assigned and the degree of structure and help provided. Weaker students may be assigned less demanding variants of a problem so that they can cope. In addition, the teacher may provide greater structure to their tasks and provide more extensive help in revising their tasks.

More able students, on the other hand, may be given greater autonomy in all aspects of their tasks. This type of differentiation is the essence of good teaching, but it reduces the standardization needed for more accurate assessment.

Koretz et al. comment:

> *"Greater standardization of tasks, revision rules, test preparation, and the like will lessen threats to validity and will probably increase scoring reliability. Such standardization, however, runs contrary to many of the basic goals of portfolio and other embedded assessment practices."* (p.14).

Group Assessment

Further tensions arise with the introduction of assessment based on co-operative small group work. Many programmes, including the APP, stress the importance of increasing the degree of group interaction and with such processes embedded in assessment, this leads to the assessment of group products. Such assessment can give distorted estimates of the competence of some students. Webb (1993) tested students on a specific type of mathematical problem – both in groups and individually. She found that group marks and individual marks corresponded reasonably for about half of the group members, but that for the other half, marks on the group test were much higher than marks on the individual test. The difference in scores was related to individual ability and the extent and type of help students took from others. In assessment based partly on group work, the raters may have no way of discerning which students' products reflect primarily the quality of other students' efforts.

It can be seen from the discussion above that there will be inevitable tensions between the activities performed by teachers in embedding assessment into their daily teaching and the standardization usually deemed necessary for good assessment. Hence, an assessment system requires practical tradeoffs arising from the degree of task standardization, the number of tasks on which assessment is based, the degree of recording and storage of information to justify a level and the impact on teachers' workloads.

Models of assuring teachers' judgments

As has been stated previously, there is a general acceptance that most teachers can rank their students reliably within a school. There would inevitably be exceptions to this, but the widespread use of teacher assessments in high stakes environments attests to this belief. This being understood, there is still a need to moderate the results of students within a school to ensure that they are comparable from class to class. In more traditional settings, this would be most validly performed by using at least some common assessment tasks.

An alternative or additional approach is for teachers to select work samples from their class that purport to be at certain performance levels and to meet with other teachers at the school to discuss the comparability of work samples with respect to the standards criteria. The latter task is seen as a way of teachers' internalizing the standards and is a valuable form of staff development. Whether the latter is sufficient by itself as a form of within-school moderation depends on how important it is to ensure comparability of the assignment of students to performance levels. When targets for the numbers of students achieving attainment levels are set then the comparability of the setting of these levels may be considered 'high stakes' and may require more secure moderation.

The second issue is moderation of results across different schools. This can take a number of forms as indicated in the APP evaluation reports. In the APP process, the moderation is based on teacher evidence collected and recorded in line with the assessment guidelines. As has been discussed previously, the issue of standardization is potentially important in the collection of evidence. Differences in the degrees to which teachers assist in eliciting evidence can have a significant effect on moderation. These differences include aspects such as structure and prompting, editorial assistance, resources provided and time allowed for the task. If part of the work is performed at home (as in an Art project), then further uncontrolled variance may be introduced through the actions of parents and friends, or even differences in the resources available at home.

The multiple purposes that are served by the assessment programme usually require some tradeoffs. The main theme of assessment for learning is generally accepted as a worthwhile goal that has the potential to enhance learning and transform children's' attitudes to learning. However, issues such as the degree to which the assessments are high or low stakes and the level of public confidence in the process are important. Further, the process must be implemented without imposing too big a workload on teachers.

One possibility to consider, which may strengthen several areas mentioned above, is to implement a few common assessment tasks as is done in the New York State and Queensland systems. The number of such tasks must be kept to a *bare minimum* otherwise, they become another burden on teachers. The tasks would be set by the central agency (the QCA) and should be able to be completed in a relatively short time, appropriate to a school's timetable. The tasks would include marking guidelines to enable them to be marked by the teacher in the school. To assist this process, annotated work samples (at different performance levels) of attempts at the task should be included. These tasks would be administered under standardized conditions. The use of such tasks would strengthen the assessment programme in a number of ways:

The moderation of assessments within a school.

Teachers would be able to refer to performances on the common task in assigning students to performance levels. The tasks would be a guide, not the final arbiter as teachers could incorporate their judgments of students' performance while working in the classroom on other non standardized activities.

The recording of evidence

Each student's results on the common tasks would be recorded and stored at the school. The teacher could make a holistic judgment as to the level of achievement of a particular student by referring to the tasks. The teacher could also draw upon further information based on classroom observation that could be used to modify the judgment of the level of achievement.

The reliability of within-school judgments

As the tasks are standardized, the results combined across (say) three such tasks would provide the necessary structure to assist in providing reliable judgments. This may help the novice teacher improve the quality of the assessments.

The validity of within-school judgments

If the tasks were well constructed, they may be used to ensure that certain parts of the curriculum were being assessed. This may guard against the possibility of aberrant teachers covering an unduly narrow curriculum that focused mainly on the teachers' interests.

Assistance in internalizing the national standards

The administration of such tasks over a period of time would help the teacher recognize the type of work that is associated with a particular performance level. This would lead to more efficient work at moderation meetings.

Improving the basis for moderation

It is likely that the administration of such tasks would produce collections of work that were more aligned to the national standards. Secondly, the teachers participating may have an improved understanding of the standards as a result of such tasks. Thirdly, the performances on the tasks themselves could be an aid in aligning the performance bands.

It should not be expected that such tasks will solve all the problems of teacher assessment, but they would provide structure, guidance and objectivity. By making it mandatory that only the results on the tasks be recorded and stored, the burden of excessive record keeping is removed from the teacher. At the same time, teachers may wish to record other aspects of learning that occur as part of the teaching process or may build up a mental picture of a student over time as a result of informal observation. These records and observations may be used to modify the results on the standardized tasks to give a fuller picture of the student.

External checks on the levels awarded

Public confidence in a system of assessment may be satisfied when it can be shown to be consistent with other external measures. Without comparisons with such external measures, the perennial issues such as "grade creep" can be a concern. It would be desirable to compare the distribution of APP performance levels with the distribution of such levels obtained on external tests. The external tests could be used in two ways:

1. To identify schools whose moderated level judgments differed markedly from their levels on the external test.
2. To check the overall levels awarded by the APP process against the levels indicated by the external tests.

Although the position of what external testing will remain is somewhat unclear, it seems that there is the capability of performing these comparisons at the end of Key Stages 1, 2 and 3.

In the NSW system, an external test is used to identify school groups whose results appear to be aberrant. Consider the assessment of achievement at the end of Key Stage 2 (Year 6). At this point, the expected level of achievement is Level 4. Suppose for the sake of simplicity, that in Year 6 the APP process produced a distribution with various percentages of students assigned to the following bands: 3c, 3b, 3a, 4c, 4b, 4a, 5c, 5b, 5a. Assume that these percentages on a national basis are appropriate, as we are only interested in school groups that are outliers. There is no questioning of the national distribution at this stage.

To identify the outlier school groups, the external test marks are grouped into levels which have the *same percentages as the APP percentages* in 3c, 3b, 3a, 4c, 4b, 4a, 5c, 5b, 5a. Then a particular school is examined to compare the school's distribution on the APP with the school's distribution on the external test. For example, if 25% of the school group is awarded 4a by the APP but only 3% are in the 4a band on the test, the school would be identified as an outlier. This identification would then allow a team of external moderators to visit the school to investigate the validity of the results.

For point 2, the overall national distribution of levels produced by the APP process is compared with an independent assessment of the levels through an equating/standard setting process on the external test. Given that these are two independent processes for assigning the percentages to the levels, approximate agreement would be expected. If for example, the APP process awarded 5% of Level 4a but the external test awarded 20%, the LEA could investigate the reasons for the difference, determine which was more appropriate and what action to take.

Cheating and authenticity of student work

The competitive pressures in education systems as well as the increased use of testing and accountability measures has led to concerns about whether or not such trends have led to more attempts to cheat. Systems which rely on external testing usually have considerable investment in the security of the tests. However there is much debate about test preparation and the ethics surrounding narrowing student experience to intentionally improve test performance.

In the USA concerns about test preparation practices of teachers have been the subject of research and discussion for many years. Mehrens & Kaminski (1989) reported more pressure to teach to the test as schools and teachers were increasingly judged by the scores their students obtained on standardised tests. They suggested that what constitutes appropriate test preparation would be under debate for some time to come, and argued that the likelihood of cheating is affected by the likelihood of getting caught, what is gained, and the consequences if caught. Commercial materials provide teachers with ways to engage students in activities to increase their scores. However these materials may be inappropriately close to the test. Although the leading school test directors understood this issue there was a lack of formal policy concerning test preparation.

Moore (1994) found that teachers were much less likely than specialists to label a practice as ‘inappropriate’ or ‘unethical’, only post-test intervention was considered to be inappropriate by teachers. Specialists considered post-test intervention; previous form preparation; during test intervention and current form intervention to be inappropriate practices. Specialists were less certain for practices reflecting motivational activities and also same format preparation. A reasonable conclusion is that teachers and test developers see the task of preparing for tests in very different ways. Moore concluded that how to inform teachers and administrators of the parameters of appropriateness of test preparation is as important as deciding how to conceptualize appropriateness.

Lai and Waltman (2008) point out that ‘even when guided by the same set of standards and guidelines, a practice deemed acceptable by one professional may not be classified as such by another’ (p29). They reported that a teacher’s determination of a given action as appropriate was not governed by consistency with professional ethics. They conclude their study of 3800 teachers from 131 schools with the warning:

> *“Our results suggest that teachers are using test-preparation practices likely to inflate test scores, rendering them less representative of students’ true achievements. When policy decisions are made on the basis of achievement trends overtime, policymakers should be aware of the factors influencing students’ test results, especially those extraneous to the construct being measured “(p.41).*

These studies indicate that codes of practice even when formalized are not without ambiguity and variation in interpretation.

In relation to APP, in schools where the teacher is the only person making assessment judgments, the extent to which these judgments are influenced by accountability pressures is a key issue. Some reports on the APP pilot data suggest that teacher judgments may be more generous than those of external moderators.

> Report 2**:** In the optional tests for year 7 (4, 5, 6) agreement between these results and ongoing teachers assessment was 48% (n=413) for reading and 57% in writing (n=383). The results that did not agree were mostly cases of overestimation by the 'ongoing teacher assessment', 43% and 32% overestimated by one level for reading and writing respectively.
>
> The optional results for year 8 were also similar with agreement at 52% (n=432) in reading and 57% (n=400) in writing, as with the year 7 results the teacher assessment level was higher in those results that did not agree.
>
> In 2004 year 9 key stage three level judgments were in agreement with ongoing teacher assessment for 61% (n=630) of pupils in reading and 70%(n=639) in writing. Where there was disagreement in writing teacher assessment was likely to be higher than the test level, for reading teacher assessment was just as likely to be higher as it was to be lower.
>
> In year 9 reading 2005 teacher judgment levels and the outcome from the national tests were the same for 56% (n=676) of pupils and for writing 59% (n=640) of pupils. When the outcome was different teacher assessment was more likely to be higher than the test level.
>
> In this data there is a clear tendency for teachers' judgment of levels to be higher than those achieved on tests. This may be because they are looking at aspects the tests are not covering. It highlights the need to ensure teacher judgments are made with an understanding of the features of performance at each level and to be satisfied that teacher assessment is standardized and moderated.
>
> (Monitoring pupils' progress in English at key stage 3. Final report on the 2003-05 pilot, February 2006.)
>
> Report 4**:** Percentage of agreement between optional test results and teacher assessment judgments for writing (n=40) was 33% and reading (n=32) was 41%. For both the test result was more likely to be higher than the assessment sub-level, however the sample was small.

Maths (n= 49) Ma1 was in agreement with test results at 29%. If there was a difference it was due to the test result being lower than the teacher judgment. Ma2 = 21%, Ma3 = 25% and Ma4 = 23% agreement, if there were differences here it was more likely that the test was higher than teacher judgment.

(Monitoring Children's Progress Project. Evaluation Report, July 2006)

Report 5: In reading moderator and school agreement ranged from 59% (AF3) to 95% (AF1),. In writing moderator and school agreement ranged from 74% (AF3, 6) to 86% (AF8). The majority of cases of disagreement were overestimations by the teacher assessment in both reading and writing.

(Evaluation of the Monitoring Children's Progress Pilot Project 2006-7. First Interim Report to the MCP Steering Group, February 2007.

Report 9: Judgments were confirmed by moderators in 67% of cases for KS2 English reading and 72% for mathematics Ma1. In cases where the level awarded for the AF was adjusted, this was most likely to involve a drop of one level.

(Report on trial of models of moderation within APP, September 2008)

In school assessment regimes where the assessment depends on classroom activities teachers bias and tendency to inflate results requires some moderation process.

Whose work does a portfolio represent?

Herman, Gearhart and Baker (1993) raise the fundamental issue as to whose work the portfolio represents. They cite other research of theirs (Gearhart, Herman and Baker, 1993) that shows that regular classroom conditions can produce considerable differences in teacher support given to students. These differences included such aspects as structure and prompting, editorial assistance, resources provided, time allowed, and so on. These differences can be problematic for the moderation of school assessments based on portfolios. If this moderation is placed in a high stakes environment, the pressures to increase instructional support in portfolio production may be substantial.

Plagiarism

Any form of submitted work which can be completed outside the classroom is potentially susceptible to plagiarism. With the advent of sophisticated search engines on the Internet the issue of plagiarism has become a major issue for assessment systems, though how much plagiarism has increased is a matter of dispute (Scanlon,2006). Student work can involve downloaded material which can distort the judgment of scholarship. Distortion can come from the fact that there is a continuum ranging from unacknowledged and fraudulent characterization of the work to careless refraining from referencing the material. The response to this has been to develop institutional frameworks to provide guidance to students about acceptable and unacceptable conduct (Park, 2004). While there are technological solutions to check authenticity of submitted essays there are always judgments to be made about what is acceptable along the continuum.

Students may seek other forms of help outside the classroom to assist in their submitted work. How much help is legitimate? At what point is the submitted work more a result of helpful assistance, than an indicator of what the student knows and can do? It is not easy to define the answer to such questions. In NSW the Board of Studies referred some matters of alleged unprofessional conduct on the part of teachers to the Independent Commission on Corruption. Using a strictly legal framework the Commission found the teachers not to be corrupt, but indicated a need for the Board to be clearer in articulating the extent of assistance which is legitimate and to educate teachers, students and parents in a code of conduct. As a result the Board now has a website and education programme to guide students, parents and teachers (http://amow.boardofstudies.nsw.edu.au/).

What works best and issues for development

The APP is an innovative approach to integrating teaching and assessment to improve student learning. It is essentially a professional capacity building programme to increase teachers' sensitivity to the developmental progression of each of their students. By emphasising the opportunities to recognise and identify signs of progress with respect to AFs in both informal and formal tasks and exercises a richer expression of the curriculum can be delivered and assessed.

As has been outlined in this review, there is a body of research that shows that considerable learning gains can be obtained when assessment for learning is employed. Implementing a system which restores teacher professional judgment to a central place in the teaching learning process is highly desirable. The complication to what should be a relatively straight forward activity of capacity building is the potential for the strong accountability agenda to undermine this activity.

In the practical implementation of such a system, tradeoffs are invariably necessary to partially achieve competing purposes. These have been discussed in the review and will be summarised here in a discussion of teacher assessment through the APP.

The reliability, validity and comparability of teachers' judgments

At the heart of the programme are the procedures by which teachers assign the performance levels at which each student is working. Ideally, these teacher decisions should be highly reliable, valid and comparable across classes within a school and across schools. The review has shown that in assessment systems similar to the APP, it is possible to gain high levels of reliability. These reliability estimates have been expressed in terms of correlations between independent markings of student work or as the percentage of perfect agreements, disagreements by 1 level, disagreements by 2 levels, and so on. However, this reliability cannot be taken for granted. Some systems have reported disappointingly low levels of reliability despite the implementation of training schemes for the assessors.

The APP uses a well structured system, with the assessment focuses being clearly described. The Assessment Guidelines also clearly describe the characteristics expected of the typical student at each attainment level. In addition, the Standards Files provide good quality work samples. Within any given school Year, the number of classifications into which students will be assigned is not large. For example, in Year 6, the expected level of attainment is level 4. Thus the majority of students in this Year will be assigned a high 4, a secure 4 or a low 4. There is evidence to suggest that for most teachers the reliability of judgments based on the APP system are satisfactory for such assignments.

A second facet of good measurement is validity – that the assessments actually measure what they purport to. In the case of APP, the assessments purport to measure a wider range of attributes (i.e. good construct representation) than would be typically measured by achievement tests. The structured APP programme, which requires teachers to engage with all the assessment focuses, is an effective way of achieving this breadth. An important test of the validity of such assessments is whether there is congruence between the actual knowledge and skills that the students possess, and the attainment levels that they have been awarded. The extent to which this is achieved can be estimated using an external judgment, for example through external moderation or by comparison with standards-based performance levels obtained on external tests. In the QCA evaluations, an examination of the overall distribution of levels awarded under APP compared with those resulting from external moderation and from optional tests showed a reassuring similarity. This suggests that the APP process will result in acceptable levels of validity when fully implemented.

A third facet of good measurement is that the levels awarded should be comparable across different classes both within a school and across schools. This has been traditionally a more difficult target to achieve for teacher assessments. From the general research literature, it would appear that assessment guidelines and standards files are not sufficient to obtain satisfactory levels of comparability – some form of moderation is usually necessary. The APP addresses this need through the planning to have a trained assessment specialist in every school. Such a specialist would be able to induct new staff in assessment practice and coordinate assessment standardisation and within school moderation.

As is indicated in the literature, concerns about the reliability and validity can sometimes create a tension between quality of measurement and good teaching practices. The former places an emphasis on standardization so that students are being compared fairly on the same or similar tasks. On the other hand, the latter often requires differentiation, where teachers may give more structure and more help to lower ability students and give greater autonomy to high ability students. In the APP process illustrative tasks are provided but there is no requirement for system-wide standardization of tasks. Assessment in APP is based on contextually valid teacher developed tasks and naturalistic observations from classroom practice, with inferences assisted by the structured guidelines.

To assist in strengthening the reliability, validity and comparability aspects of assessment, it is therefore suggested that two possible models are available. The first, used in Scotland, takes a sample survey of student performance in specific year cohorts, and feeds national data back into the schools to enable them to self-evaluate their standards of achievement against the national data.

The teachers' judgments are supported by exemplar student work materials on the National Assessment 5-14 Bank. However, the system is vulnerable to patchy uptake. The preferred model arising from this review is that some externally developed standardized performance tasks should be introduced.

These common assessment tasks would be similar to those in the New York State education system and the Queensland education system, both of which emerged from the struggle to maximize reliability, validity and comparability with as little disruption to teaching and learning goals as possible. Such tasks could be developed centrally by QCA to ensure they are easily administered as a natural part of classwork.

Note, however, that the literature suggests that if student performance results from either the external standardized tasks or the teachers' assessments are intended for summative reporting at a discrete school level, the schools will try to maximize performances and there will be negative impacts on teaching and learning as a result. Therefore to avoid any detrimental backwash on classroom teaching and learning processes, care would need to be taken to ensure that the assessments would not be perceived as high stakes by either the school or the students. They would be marked in school by the class teacher according to the marking guidelines and annotated work samples provided by the QCA.

The number of such tasks should be kept to a minimum but should be sufficient to provide some standardized information of use to teachers in assessing the level at which students are working. They would provide part data in assessing each student. The teacher would make a holistic judgment on the basis of such tasks and other data that the teacher has collected or observed. A below expected performance on a task by a particular student could be overridden by the teacher if other observations showed that the student was generally working at a higher level than indicated by the common task.

If common assessment tasks are systematically used, the potential advantages are the greater standardisation of the basis for making teacher judgments and the support they would give to inexperienced teachers. The downside is the potential for the tasks to become the main focus of assessment and for busy teachers to unduly rely on them. For some teachers this could possibly narrow the range of assessment.

Issue for consideration: common assessment tasks

That consideration is given to strengthening the reliability, validity and comparability of the APP assessments by the use of common assessment tasks that could be administered naturally as a part of classroom assessment.

The range of work and manageability issues

Teachers use the Assessment Guidelines to assess each pupil on the assessment focuses and a flowchart to assign an overall level of working to each student. The efforts to improve the validity of assessments by basing them on a wide range of evidence can create workload issues for teachers. In the QCA evaluation reports, the teachers indicated that finding the evidence to support their judgments was the most challenging part of the process. The management of the storage and retrieval of such evidence is an additional burden. Similar systems overseas have reported workload issues where teachers have felt under pressure to manage the data collection and storage.

Workload and evidence gathering have remained persistent challenges in teacher assessment contexts since the early days of the National Curriculum. During the first few years of the reform, assessment was highly structured and based on attainment targets, statements of attainment and a 10-level scale of achievement. According to Wilmut (2004) schools felt there was insufficient central guidance and they struggled with concerns about workload, the time taken to carry out assessments with normal class sizes (and its detrimental impact on teaching/learning time) and the nature and volume of student work that had to be retained to provide evidence for the assessment judgments made.

When the system is implemented and stabilized, if teachers report that the workload is excessive, then consideration should be given to ways of simplifying the amount of evidence that is mandatory to record and store. Teachers should have the freedom to record and store as much evidence as they can comfortably gather in their daily classroom activities. What is being suggested here is that a *minimum* amount of data be formally collected and stored, perhaps less than the four to six pieces of writing from different subjects, reported in the APP primary literacy pilots. If the recommendation to introduce common assessment tasks is accepted, then it is suggested that the scores on these tasks be retained for a given period in case a student challenges their attainment level.

There is an obvious trade-off concerning teacher workload and the amount of record keeping. In the early years of school, concerns are not usually raised concerning the issue of formal record keeping. However for the later years, when students are moving towards qualifications with higher stakes, students and their parents may be more inclined to challenge the performance level. Under such challenges, it is advantageous for a teacher to be able to point to clear indicators of achievement that have been formally recorded.

Issue for consideration: that record keeping is kept manageable

That care should be taken in both the APP design and its implementation processes to ensure that record keeping is manageable. Advice should be given on continuous refinement of assessment activities to ensure that workload is not burdensome, that time is built into the daily/weekly system to enable adequate recording and reporting, and that appropriate technology (e.g. hand-held devices) is used to increase administrative efficiency.

If common assessment tasks are introduced, then scores obtained on them should be stored for a suitable time to be used in evidence for various purposes. This information would be supplemented by any additional information that teachers choose to collect.

Models of assuring teachers' judgments

The APP programme has well-developed advice and processes for both within-school and across school moderation, using a choice of modes for the latter. Evidence from the pilot studies shows that teachers initially differ in their subject knowledge, assessment practice and pedagogical understanding. These moderation processes have been reported as being useful for developing teacher assessment capacity and their understanding of what is required for the attainment of each level.

However, if external testing is available at the end of the key stages, then it would re-assure public confidence if the results of moderation were consistent with the results from external testing. This could be performed in two ways. Firstly, at the school level, the external tests could be used to identify outlier school subject groups that award too many or too few students in a particular level than would be expected on the results of the test.

This analysis would accept the percentage of students at each level on a national basis (as awarded by the APP programme) as a given. The parameters of the identification could be set to control how aberrant a school group's results on the tests could be before being flagged by the comparability algorithm. The identified schools could then be invited to investigate the difference(s), with a reporting mechanism for actions planned to remediate the situation as appropriate, or they could be visited by a moderating team to discuss possible reasons for the discrepancy and contribute to professional development activities. In either case there should be discussion on the actions to be taken to ensure a more accurate representation of the students' levels of achievement.

At the national level, a second procedure would be to use both types of assessment by checking the level distributions from the APP against the level distributions resulting from an external test. Naturally, one would not expect perfect agreement here – just as one finds differences when two different standard-setting methods (e.g. Angoff multi-stage and Bookmarking) are applied to the same data. However, too radical a disagreement would prompt an analysis of reasons for the difference. In addition, if one measure shows a consistent trend of change over time and the other shows stability over time, then this makes it very difficult to interpret the pattern of results.

Should testing at the end of key stages and/or single level testing not continue, then any major shifts in the school assignment of levels should be monitored and 'evidence checks' be instituted so that any 'drift' in judgments without firm evidence is countered.

Issue for consideration: periodic external checks

That schools should be supported in developing the accuracy and comparability of their judgments through external processes designed to monitor the levels awarded by their APP system. If any outlier school subject groups are identified, the expectation should be that the schools will investigate the reasons for any major discrepancies and take appropriate action.

Issues of cheating, authentication, appealing against results

It is important to address the fact that any assessment regime has to be embedded in a professional framework with specification of the ethical perspectives relevant to the system employed. Outright cheating is easier to deal with than the more subtle influences which can introduce bias into the assessment process. As the literature suggests, there are often discrepancies between the views of teachers and assessment professionals about the appropriateness of certain preparation practices. Despite advice to the contrary, for example, teachers in many education systems have been found to narrow their delivery of the curriculum to those elements that figure prominently in the assessment regime. The fact that assessment is used for so many purposes and can be based on different types of evidence means that what constitutes good professional practice needs to be clearly documented.

In the APP system which clearly documents the AFs to be covered there should be little evidence of curriculum narrowing if the protocol is followed. However the fact that APP encourages use of classroom observation and informal observations based on a range of classroom interactions means that the evidence base needs to be seen as fair and appropriate for all students.

There is a range of issues relating to the resolution of the observer/facilitator nexus that can become an issue in the formative focus (which should be the major emphasis in assessment for learning), and the education system reporting function, which requires reliable and valid reference to expected standards. For example, depending on the stakes attached to the process of teacher assessment there needs to be clear expectations about the evidence base for reporting at different points of feedback, ranging from feedback on a given assignment to less formal feedback given in various classroom interactions. Lack of clarity can lead to students and/or parents wanting to appeal what they may consider invalid or inconsistent judgments.

Concerns about the extent to which there will be on-going issues about authentication and whether or not appeals against judgments will be made will depend upon the extent to which there is confidence in the extent to which schools properly implement the programme. Most education systems find consistent deployment in the implementation of new requirements or processes can be problematic. As has been pointed out teachers can make ‘local’ adjustments to processes without realizing the consequences for the success of the programme.

Issue for consideration: a code of professional practice

That a code of professional practice be included as part of the APP programme and that appropriate forms of the code are made available to teachers, students and parents.

The impacts of an increased focus on teachers' professional judgments on teaching and learning and the curriculum

Evidence from education systems where teacher assessment has been implemented with major professional support, is that everyone benefits. Teachers become more confident, students obtain more focused and immediate feedback, and learning gains can be measured. An important aspect of teacher assessment is that it allows for the better integration of professional judgment into the design of more authentic and substantial learning contexts.

Familiarity with and confidence in using APP should lead to a more integrated approach to teaching, learning and assessment. In contrast, imported test and worksheet focused assessments have convenience but often are not appropriately adapted to particular classroom contexts of integrated learning. Hence adoption of APP should lead to questioning of materials and practices which are not consistent with this approach. It will give schools occasion to review the role of additional testing or assessment practices which may be currently in place.

Issue for consideration: periodic review of assessment programme

That schools should be advised to review all school assessment practices to ensure their consistency with APP. The use of externally sourced tests and worksheets, which might have featured prior to the adoption of APP, should be reviewed.

The use of teacher judgments in periodic reporting to parents

Teacher judgments in periodic reporting to parents within a common standards framework have become routine practice not just in the UK but across the world including, for example, the Australian States covered by this review. In Australia, the initial systems tended to be over-specified and detailed, resulting in a national policy that adopted plain language reporting to ensure that students and their parents could see their progress within a common reporting format. To accomplish this task, work samples and professional development programmes were developed. The reporting function is therefore considered to be one of the most important features of any assessment system and no less so in an APP context. In addition to its formative and diagnostic purposes, APP should therefore assist teachers in providing students and parents with accessible, meaningful and accurate assessments of progress.

Issue for consideration: development of common reporting format

That schools and teachers should be facilitated, through professional development opportunities and resource materials, in developing reporting processes that use a common format to provide accessible, meaningful and accurate assessment of progress to students and parents.

Continuous professional development requirements

A systematic programme of providing ongoing school-based professional development should be part of the APP system. This process should build both assessment expertise and comparability of assessments across classes, and a culture of integrating assessment routinely into classroom processes of teaching and learning.

To enable personal professional development, referencing and calibration there needs to be a good supply of student work exemplars, electronically available and regularly augmented and refreshed. In addition, teachers need to be trained in the application of APP materials through both in-house training programmes and training by external assessment experts. This training needs to be practical, by allowing teachers to actively engage with the APP materials in a real world context. If possible, the training should incorporate teachers applying the APP approach to their own students' work in their own classrooms.

Teachers need to be given the opportunity to work with more experienced colleagues in their own school and with colleagues in other schools to practice the APP principles. Time needs to be scheduled to allow this to occur. It is this practical approach that will enhance their sense of professionalism, motivate them and provide their commitment to the programme. Such an approach will require the assistance of senior leaders with the school and pro-active support by the LAs to help embed the assessment practices.

Professional development clearly emerged as an important factor in the pilot studies. Even with the excellent range of support materials now available in the formal roll-out of APP, there should be no less investment in professional development and teacher capacity-building.

Issue for consideration: programme of professional development

A systematic programme of school-based professional development should continue to be built into the APP system. This should be facilitated through centrally designed and provided resources to help ensure consistency of message and practice.

Considerations for an Evaluation of the APP system

Once the APP system has been operating in schools for a few years, it is important that it be evaluated. In the initial implementation, as teachers and school/local area administrators engage with the issues of applying a system as potentially useful as the APP, there may well be "teething problems" . In other systems it has been found that it may take up to three years before the system reaches stability in implementation. In this time, information about the strengths and weaknesses of the system would be gathered informally, which could be used to fine tune the implementation. However, after three years it is suggested that a more formal evaluation take place.

What the Evaluation could comprise

An important part of any evaluation is to describe the programme implementation – how it functions in practice. A programme may articulate certain principles and practices as the core of its ethos, but the implementation may have a different emphasis. A key aim of the evaluation could be to describe what typically happens in a classroom when APP is implemented. That is, in practice how do a teacher's classroom actions using APP differ from a teacher who is not implementing APP. Some pertinent issues are:

- How is information on student performance actually captured in practice?
- What proportion of data is stored and how is it stored?
- Are the student assessment activities evenly spread across students?
- Does the assessment gathering change the classroom dynamics?
- Are some assessment focuses captured more readily than others?
- Are teachers assessing more broadly than previously?
- Has there been a change in teachers' workloads?
- Are teachers' assessments more accurately aligned to the national standards?
- Has reliance on formal testing decreased?
- Have there been changes in the classroom climate through APP?

The APP programme has a strong focus on the validity of assessment which gains its breadth through attending to the assessment focuses. It is important to find out what effect the implementation has on teachers' workloads. If the workloads have increased, how have teachers adapted the programme to cope in practice? The range of practices that teachers use to gather assessments could be described and analysed in relation to the workloads required and the effectiveness of the practices.

A further important issue is the extent to which teacher assessment has shifted away from the use of formal testing with the implementation of APP. Has the latter decreased with the implementation of APP or has it remained at former levels?

A major aim of APP is the development of capacity-building in teacher assessment. It would be useful to determine whether, after APP implementation, teachers' accuracy in assigning pupils to national performance levels improved over their baseline levels. A related issue concerns the extent and type of moderation practices. Under APP, what moderation practices are typically used within a school and between schools to gain comparability of performance levels?

Teachers could also be surveyed about the affective characteristics of the implementation of APP – whether it has enhanced their image of professionalism and sense of collegiality within the profession. It may also be possible to obtain student input on improvements in the learning environment in the classroom under APP.

Gathering information for the Evaluation

There are several models that could be used for the evaluation, that are not mutually exclusive. One model would involve gathering information on baseline measures within a school before the implementation of APP, allowing APP to run for some years, and gathering post APP implementation data using the same measures. These measures could be rating scales on various dimensions made by teachers, perhaps some students, and administrators. In the Campbell and Stanley (1966) quasi-experimental design, this model involves a pretest, treatment (the APP), and posttest. It has the merit that the pretests and posttests may be based on different samples of respondents, the posttest judgments being made without knowing the rating levels on the pretests.

A second model, more usually employed, involves a single set of respondents estimating (through rating scales) whether any change has occurred. This requires that the respondents had experienced conditions prior to APP, and are able to assess whether certain aspects are better, worse or unchanged after APP.

In the second model, information is often gathered in phases, where initial questionnaires identify issues that can be targeted in more depth by follow-up questionnaires and structured interviews.

Phase 1 of teacher information gathering

An evaluation of the APP system should place great weight on the views of the personnel that are crucial to its success – the classroom teachers. The teachers' views should be collected by a well designed questionnaire that gathers the essential information required with a minimum response time. Efforts should be directed to gathering as high a response rate as possible by periodic but gentle reminders. Many of the questions could be five or seven point Likert-type scales which are easy and quick for the teachers to fill in. This format would allow the computer scanning of the sheets and thus facilitate quick data analysis. There are many statistical analysis packages that are useful for such purposes; e.g. SPSS or SAS. Space for teachers to write their comments should be given and these comments should be read and analysed on a qualitative basis. Such comments are often extremely valuable as the fixed questions may be structured in such a way as to truncate issues that are important to teachers.

Phase 2 of teacher information gathering

Depending on the initial questionnaire responses, follow-up questionnaires could be given to target issues that arose from the first questionnaire. These could be quickly analysed by computer scanning and statistical analysis through packages such as SPSS. These could provide relatively large numbers of responses that effectively sample the population of teachers in the implementation.

However, in this second phase, it is important to conduct structured interviews with teachers to obtain more accurate and detailed information. From the first phase, schools can be identified that are successful APP implementers as judged by their positive responses to the written questionnaires. Similarly, schools that are negative to (or struggling with) APP can be identified. Structured interviews with the staff of such schools may be able to elicit reasons why the programme is succeeding in some schools and struggling in others. It may be also the case that some school subjects may be more amenable to the APP approach that others. For example, English teachers may have different views on the implementation of APP to Mathematics teachers.

The information obtained from teachers could be targeted to reflect the role of the teacher in the school. Different questionnaires could be given to senior staff whose roles were primarily administrative than to teachers whose work was primarily classroom based.

Similarly, staff in local areas could be given different questionnaires reflecting any changes in workloads or costs in administering the programme compared to a previous baseline.

Some pertinent issues for the teachers could include the effects on their workload of gathering the APP information and the effect that the information gathering is having on the classroom interactions. An important part of the information gathering is to establish *how* the programme is being implemented. In such a programme, there may be some variation in the implementation. As is seen in the literature, teachers may modify such programmes to suit their own cognitive styles and classroom procedures. Such variations are an inevitable part of implementing a new programme.

Identifying positive and negative implementations

An important aspect of the evaluation could be to correlate such different classroom practices with successful implementation of the programme. A number of criteria could be used to measure successful implementation – important ones would be positive teacher attitudes to the programme and positive attitudes from the children in the class as measured by questionnaires and structured interviews. Such criteria could also include the success to which such teachers are able to judge the national standards, as determined by external checks such as common test results or moderator visits.

Ideally, one would also examine improvements in learning from the students as a result of the implementation of the programme. Various criteria could be used to identify successful implementation and unsuccessful implementation. Having done this, one could examine differences in the actual way APP was implemented in the two groups. This could give valuable insights into ways to improve the programme.

External Validation of the Measures

In addition to measuring success at the school level, it would be necessary to assure other stakeholders that the programme is achieving high quality measurement that is reliable, valid and comparable across schools. At suitable times in the programme, annotated student work samples could be gathered and assigned a national achievement level by the teachers. These could then be marked by external teachers and/or by external moderators to check whether the school assigned levels were appropriate. These measures would allow reliability estimates of the work sample marking to be calculated, both in terms of correlation coefficients and extent of misclassification of levels.

In addition such measurements would enable the checking of the extent of the comparability of the levels. Such checks should cover a reasonably full range of the achievement distribution to ensure comparability does not vary across this distribution. For example, a school may have suitable comparability for the high achievement levels but may have poor comparability for the low achievement levels. In other systems (for example NSW), it has been shown that some teachers may over-estimate the performance levels of low ability students while being accurate with the performance levels of high ability students.

Evaluation of levels over time

In order to retain public confidence in the levels awarded, such evaluations should periodically check the distribution of levels awarded over time to determine whether there is any systematic change. This checking could look at local area results and national results. If it turns out that a particular local area is consistently improving its results, an examination of the reasons for this success would be useful.

From one year to the next, fluctuations in levels awarded can develop, so that observations over some years are required in order to detect trends. If trends are detected (for example, percentages in the higher levels are increasing), then this potentially positive result should be checked against other external measures.

For example, the administration of stratified randomly-parallel tests to samples of schools at different time intervals could allow inferences to be drawn about whether high level results were increasing in that sample. Any inferences would have to take account of the estimated standard error of measurement and the degree of parallelism of the tests. This data could then be compared with the distribution of levels over time assigned by APP in the same sample.

Evaluation of affective aspects on teacher development

One of the most important potential results of the APP programme could be an enhancement of the teachers' sense of professionalism which could lead to greater commitment and improved classroom practice. This could also be measured by a short written questionnaire comprising Likert scale type items with open-ended sections for teachers to comment. Follow-up interviews with teachers could enable the evaluators to probe responses at a greater depth. The data obtained on the affective aspects of teacher development and a sense of increased professional enhancement may have flow-on effects on teaching. Given the data gathered by the questionnaires, and the distribution of performance levels over time, it may be possible to test this hypothesis.

Glossary of Acronyms

ARC	Assessment Resource Centre
AF	Assessment Focus
APP	Assessing Pupils Progress
ENTER	Equivalent National Tertiary Entrance Rank
ERIC	Educational Resources Information Centre
GAT	General Achievement Test
GCSE	General Certificate of Secondary Education
HSC	Higher School Certificate
ICT	Information and Communications Technology
IRT	Item Response Theory
ITBS	Iowa Tests of Basic Skills
KR-20	Kuder-Richardson 20 coefficient
LOTE	Language Other Than English
NAA	National Assessment Agency
NSW	New South Wales
OAI	Overall Achievement Indicator
OP	Overall Position
QCA	Qualifications and Curriculum Authority
QCAR	Queensland Curriculum, Assessment and Reporting
QCATs	Queensland Comparable Assessment Tasks
QCE	Queensland Certificate of Education
QCS	Queensland Core Skills
QSA	Queensland Studies Authority
RUMM	Rasch Unidemensional Measurement Model
SAI	Subject Achievement Indicator
SC	School Certificate
SSA	Scottish Survey of Achievement
TA	Teacher Assessment
UAC	University Admission Centre
UAI	University Admissions Index
VCAA	Victorian Curriculum and Assessment Authority
VCE	Victorian Certificate of Education
VELS	Victorian Essential Learning Standards
VTAC	Victorian Tertiary Admissions Centre

References

Airasian, P. W., Kellaghan, T., Madaus, G. F., & Pedulla, J. J. (1977). Proportion and direction of teacher rating changes of pupils' progress attributable to standardized test information. Journal of Educational Psychology, 69(6), 702-709.

American Psychological Association (1966). Standards for educational and psychological tests and manuals. Washington, D.C.: APA.

Andrich, D., Sheridan, B.S. and Luo, G. (2005). RUMM2020: Rasch Unidimensional Models for Measurement. RUMM Laboratory: Perth, Western Australia.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement. (2nd ed.). Washington, D.C.: American Council on Education.

Baker, E. L., O’Neil, H. P. and Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. American Psychologist, 48(12) 1210-1218.

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. Educational Psychology, 21(2), 177-187.

Bentley, T. (2002). What learning needs, towards educational transformation: a challenge of nations, communities and learners. A keynote paper presented at the Curriculum Corporation Conference, 27 May, 2002. Canberra, Australia.

Black, P. and Wiliam, D. (1998a). Assessment and classroom learning. Assessment in Education, 5, 7-74.

Black, P. and Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. London: King’s College London School of Education.

Boaler, J. (2002) Experiencing school mathematics: traditional and reform approaches to teaching and their impact on student learning. Mahwah, NJ., Lawrence Erlbaum Associates.

Board of Studies NSW (2002). Studying for the NSW Higher School Certificate. An information booklet for Year 10 students, 2002. Sydney, Australia: Board of Studies, NSW.

Board of Studies NSW (2003). HSC assessment in a standards-referenced framework: a guide to best practice. Sydney, Australia: Board of Studies, NSW.

Bobis, J. (1997). Report of the Evaluation of the Count Me In Too Project. Report submitted to the NSW Department of Education and Training.

Bond, T., & Caust, M. (2005). Silk purses from sows' ears? Making measures for teacher judgements. Paper presented at the AARE Conference. Sydney, 2005.

Braun, H. I. and Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Brookhart, S. M. (2005). The quality of local district assessments used in Nebraska's school-based teaher-led assessment and reporting system (STARS). Educational Measurement: Issues and Practice, 24(2), 14-21.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. Assessing Writing, 9, 105-121.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Campbell, D. T., and Stanley, J. C. (1966). Experimental and Quasi-experimental Designs for Research. Chicago: Rand McNally.

Clare, L., & Aschbacher, P. R. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. Educational Assessment, 7(1), 39-59.

Clarke, S., & Gipps, C. (2000). The role of teachers in teacher assessment in England 1996-1998. Evaluation and Research in Education, 14(1), 38-52.

Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. Journal of Educational Psychology, 78(2), 141-146.

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. Educational Research and Evaluation, 13(5), 401-434.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. Review of Educational Research, 58(4), 438-481

Cumming, J. J. and Maxwell, G. S. (2004). Assessment in Australian schools: current practice and trends. Assessment in Education, 11, 89-108.

Daugherty, R. (2009) Personal communication from Professor Richard Daugherty, Chair of the Daugherty Assessment Review Group, Wales 2003-2004

Daugherty, R. (2004) Learning Pathways through Statutory Assessment: Key Stages 2 and 3, Cardiff: Welsh Assembly Government, Department for Children, Education, Lifelong Learning and Skills http://www.amdro.org.uk/eng/Learning/Assessment/Daughety_Final_Report.pdf

Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. School Psychology Quarterly, 13(1), 8-24.

Eaves, R. C., Williams, P., Winchester, K., & Darch, C. (1994). Using teacher judgment and IQ to estimate reading and mathematics achievement in a remedial-reading program. Psychology in the Schools, 31, 261-272.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinman, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. Psychology in the Schools, 43(3), 247-265.

Egan, O., & Archer, P. (1985). The accuracy of teachers' ratings of ability: A regression model. American Educational Research Journal, 22(1), 25-34.

Elley, W. B. and Livingstone, I. D. (1972). External Examinations and Internal Assessments. NZCER, Wellington.

Falk, B., Ort, S. W., & Moirs, K. (2007). Keeping the focus on the child: Supporting and reporting on teaching and learning with a classroom-based performance assessment system. Educational Assessment, 12(1), 47-75.

Farr, R., & Roelke, P. (1971). Measuring subskills of reading: Intercorrelations between standardized reading tests, teachers' ratings, and reading specialists' ratings. Journal of Educational Measurement, 8(1), 27-32.

Feldt, L. S., and Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed., 105-146). New York: Macmillan.

Fitzpatrick, R. and Morrison, E. J. (1971). Performance and product evaluation. In R.L. Thorndike (Ed.), Educational Measurement. (2nd ed.). Washington, D.C.: American Council on Education.

Fredericksen, J. R. and Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9) 27-32.

Fuller, M. L. (2000). Teacher judgment as formative and predictive assessment of student performance on Ohio's fourth and sixth grade proficiency tests. Paper presented at the annual meeting of the American Educational Research Association. New Jersey, 2000.

Gearhart, M., Herman, J.L., Baker, E.L., & Whittaker, A.K. (1993). *Whose work is it? A question for the validity of large-scale portfolio assessment* (CSE Tech. Rep. No. 363). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Greenall, P. D. (1949). The concept of equivalent scores in similar tests. British Journal of Psychology (Statistical Section), II, 30-40.

Gipp, C. (1995). What do we mean by equity in relation to assessment? Assessment in Education 2(3) 271-282.

Harlen, W. and Deakin-Crick, R. (2003) Testing and motivation for learning. Assessment in Education 10(2), 169-208.

Hargreaves, D. J., Galton, M. J., & Robinson, S. (1996). Teachers' assessments of primary children's classroom work in the creative arts. Educational Research, 38(2), 199-211.

Harlen, W. (2005). Trusting teachers’ judgement: research evidence of the reliability and validity of teachers’ assessment used for summative purposes. Research papers in education, 20, (3), 245-270.

Hattie, J. & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77 (1), 81-112.

Hay, P. J. & Macdonald, D. (2008). Misappropriations of criteria and standards-referenced assessment in a performance-based subject. Assessment in Education, 15, 153-168.

Hayes (2009), personal communication

Helmke, A., & Schrader, F-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. Teacher and Teacher Education, 3(2), 91-98.

Herman, J. L., Gearhart, M., & Baker, E. L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. Educational Assessment, 1(3), 201-224.

Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgements of pupil achievement levels. Journal of Educational Psychology, 76(5), 777-781.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. Review of Educational Research, 59(3), 297-313.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content are using teachers' ratings as criteria. Journal of Educational Measurement, 22(3), 177-182.

Howard, M. (1958). The conversion of scores to a uniform scale. British Journal of Statistical Psychology, XI, 199-207.

Johnson, S. and Munro, L. (2008) Teacher judgement and test results: should teacher and tests agree? Paper presented at the Annual Conference of the Association for Educational Assessment – Europe Hissar, Bulgaria http://www.aea-europe.net/userfiles/13_Sandra%20Johnson_Johnson_%20Munro.pdf

Jordan, P. (2008). Externally moderated school-based assessment in Queensland – How do we know that it works? Paper delivered at the QSA Senior Schooling Conference, 10-11 March 2008.

Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England. Assessment in Education, 2(2) 145-164.

Klenowski, V. (2007). Evaluation of the effectiveness of the consensus-based standards validation process. Brisbane,Qld, Department of Education, Training and the Arts.

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and uncorrelated error on Coefficient alpha. Applied Psychological Measurement, 21, 337-348.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). The reliability of scores from the 1992 Vermont portfolio assessment program, interim report. National Center for Research on Evaluation, Standards and Student Testing.

Koretz, D., Stecher, B, Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. Educational Measurement: Issues and Practice, 13(3), 5-16.

Kuder, G. F. and Richardson, M. W. (1937). The theory and estimation of test reliability. Psychometrika, 2, 151-160.

Lai, E. R., & Waltman, K. (2008). Test preparation: examining teacher perceptions and practices. *Educational Measurement: Issues and Practice, 27(2),* 28-45.

Lindström, L. (2007). Assessing Craft and Design: Conceptions of Expertise in Education and Work. In: A. Havnes & L. McDowell (Ed.) Balancing Dilemmas in Assessment and Learning in Contemporary Education. London: Routledge.

Linn, R. L. (1966). Grade adjustments for prediction of academic performance. Journal of Educational Measurement, http://www.informaworld.com/smpp/title~content=t775653631~db=all~tab=issueslist~branches=6 - v63, pp. 313-329.

Linn, R. L. (1993). Linking Results of Distinct Assessments. Applied Measurement in Education, http://www.informaworld.com/smpp/title~content=t775653631~db=all~tab=issueslist~branches=6 - v66, pp. 83-102.

Linn, R. L., Kiplinger, V. L., Chapman, C. W., & LeMaheiu, P. G. (1992). Cross-state comparability of judgments of student writing: Results from the new standards project. Applied Measurement in Education, 5(2), 89-110.

Lissitz, R.W. & Samuelson,K. (2007). A suggested change in terminology and emphasis regarding validity and education. Educational Researcher, 36, (8), 437-448.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Matsumura, L. C., Garnier, H., Pascal, J., & Valdes, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. Educational Assessment, 8(3), 207-229.

MacCann, R. G. (1996). The Moderation of Higher School Certificate Assessments using a Quadratic Polynomial Transformation: a Technical Paper. Sydney: NSW Board of Studies.

MacCann, R. G. (1998). The relationship between school assessments and HSC marks. In J. S. Cook (Ed.), A Review of the Higher School Certificate Assessment Program. (12-24). Sydney: NSW Board of Studies.

MacCann, R. G. (2004). Reliability as a function of the number of item options derived from the ‘knowledge or random guessing’ model. Psychometrika, 69, 147-157.

McCurry, D. and MacKenzie, M. (2006). Teachers Making Contextualised, GroupJudgements of Generic Skills and Dispositions. Paper presented at the 32nd IAEA Conference, Singapore, 21-26 May, 2006.

McDonald, B. and Boud, D. (2003). The impact of self assessment on achievement: the effects of self assessment training on performance in external examinations. Assessment in Education, 10(2) 209-220

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. Educational Measurement: Issues and Practice, 20(1), 20-32.

Marzano, R. J., Pickering, R. J. and McTighe, J. (1983). Assessing student outcomes: Performance assessment using the dimensions of learning model. Alexandria, VA: Association for Supervision and Curriculum Dimension.

Masters, G. N. and McBryde, B. (1994). An investigation of the comparability of teachers' assessment of student folios. Brisbane: Queensland Tertiary Entrance Procedures Authority.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: fruitful, fruitless, or fraudulent?. Educational Measurement: Issues and Practice, 8(1), 14-22.

Meisels, S. J., Liaw, F, Dorfman, A., & Nelson, R. F. (1995). The work sampling system: Reliability and validity of a performance assessment for young children. Early Childhood Research Quarterly, 10, 277-296.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, J., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. American Educational Research Journal, 38(1), 73-95.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., 13-103). New York: Macmillan.

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. American Psychologist, 50, 741-749.

Mitzel, H. C., Lewis, D. M., Patz, R.J. & Green, D. R. (2001). The bookmark procedure: psychological perspectives. In G. J. Cizek (Ed.), Setting performance standards (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.

Moore, W. P. (1994). Appropriate test preparation: can we reach a consensus?. Educational Assessment*, 2(1),* 51-68.

Morrison, K. and Tang Fun Hei, J. (2002). Testing to destruction: a problem in a small state. Assessment in Education, 9(3), 289-317.

Moss, P. A. (1994). Can there be validity without reliability? Educational Researcher, 23(2), 5-12.

Munro, L., & Johnson, S. (2008). Exploring the validity of judgements of pupils' attainments. Paper presented at the annual conference of the International Association for Educational Assessment, Cambridge, 2008.

Murphy, R. J. L. (1979). Teacher assessments and GCE results compared. Educational Research, 22(1), 54-59.

Natriello, G. (1987) The impact of evaluation processes on students. Educational psychologist, 22(2), 155-175.

Newman, F. M., Bryk, A. S. and Nagaoka, J. K. (2001). Authentic intellectual work and standardized tests: conflict or coexistence? Chicago, IL., Consortium on Chicago School Research.

Novick, M.R. and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. Psychometrika, 32, 1-13.

Park, Chris (2004) 'Rebels without a clause: towards an institutional framework for dealing with plagiarism by students', Journal of Further and Higher Education, 28:3, 291 — 306

Pedulla, J. J., Airasian, P. W., & Madaus, G. F. (1980). Do teacher ratings and standardized test results of students yield the same information? American Educational Research Journal, 17(3), 303-307.

Phelps, R.P. (2008). The role and importance of standardized testing in the world of teaching and training. Third Education Group Review/Essays, 4 (3). Retrieved 19 Feb 2009 from http://www.thirdeducationgroup.org/Review/Essays/v4n3.htm

Pilliner, A. E. G. (1958). The rescaling of teachers' estimates. British Journal of Statistical Psychology, XI, 191-197.

Polly, C., Rahman, A. A., Rita, L., Yun, Y. D., & Ping, L. Y. (2008). An investigation of reliability and validity: using rubric approach in learning & teaching. Paper presented at the annual conference of the International Association for Educational Assessment, Cambridge, 2008.

Popham, W. J. (1987). The merits of measurement-driven instruction. Phi Delta Kappan, May, 679-682.

Queensland Studies Authority (2005). Moderation Processes for Senior Certification. Brisbane: Queensland Government Printer.

Radford ,W.C. (1970). Public examinations for Queensland secondary school students: report of the committee appointed to review the system of public examinations for Queensland secondary school students and to make recommendations for the assessment of student achievements. Dept. of Education , Queensland.

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. British Journal of Mathematical and Statistical Psychology, 54, 315-323.

Raykov, T. (2008). Alpha if item deleted: a note on loss of criterion validity in scale development if maximizing coefficient alpha. British Journal of Mathematical and Statistical Psychology, 61, 275-285.

Reeves, D. J., Boyle, W. F., & Christie, T. (2001). The relationship between teacher assessment and pupil attainments in standard test tasks at key stage 2, 1996-1998. British Educational Research Journal, 27(2), 141-160.

Roschewski, P. (2004). History and background of Nebraska's school-based teacher-led assessment and reporting system (STARS). Educational Measurement: Issues and Practice, 23(2), 9-11.

Scanlon, P. (2006). Solving the Problem of Internet Plagiarism? The Technological Expediency of Online Plagiarism-Checkers. Teachers College Record, Date Published: October 18, 2006 http://www.tcrecord.org/Home.asp ID Number: 12797, Date Accessed: 11/12/2006 4:08:34 PM

Scottish Government (2006) Scottish Survey of Achievement: Practitioner's Report and Technical Annex: 2005 Scottish Survey of Achievement (SSA) English Language and Core Skills.
http://www.scotland.gov.uk/Publications/2006/06/29141936/0

Scottish Government (2007) Scottish Survey of Achievement 2006 Social Subjects (Enquiry Skills) and Core Skills – Supporting Evidence
http://www.scotland.gov.uk/Resource/Doc/195029/0052389.pdf

Scottish Government (2008) Scottish Survey of Achievement 2006 Social Science, Science Literacy and Core Skills – Supporting Evidence
http://www.scotland.gov.uk/Resource/Doc/1038/0061218.pdf

Shapley, K. S., & Bush, M. J. (1999). Developing a valid and reliable portfolio assessment in the primary grades: Building on practical experience. Applied measurement in Education, 12(2), 111-132.

Shapley, K. S. & Pinto, M. F. (1995). Final report of the development of the 1994-95 Chapter 1 Portfolio assessment (REIS95-274-2). Dallas, TX: Dallas Public Schools, Division of Research, Planning and Evaluation.

Sharpley, C. F., & Edgar, E. (2006). Teachers' ratings vs standardized tests: An empirical investigation of agreement between two indices of achievement. Psychology in the Schools, 23(1), 106-111.

Sinatra, R., & Venezia, J. (1986). Establishing interrater reliability for kindergarten and primary-grade writers. Early Child Development and Care, 24(1), 91-112.

Spearman, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 271-295.

SQA (2007) Scottish Survey of Achievement (Science, Science Literacy and Core Skills) 2007: Guidance for Participating Schools, Booklet 1. Glasgow: Scottish Qualifications Authority www.sqa.org.uk

SQA (2006) Scottish Survey of Achievement 2006, Social Subjects (Enquiry Skills): Important Information for Teachers: Advice on Collecting Extended Writing Glasgow: Scottish Qualifications Authority www.sqa.org.uk

Starch, D. (1913) Reliability and distribution of grades. *Science*, 38, 983,630-636.

Stokking, K., Van der Schaaf, M., Jaspers, J & Erkens, G. (2004). Teachers' assessment of students' research skills. *British Educational Research Journal 30* (1), 93-116.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. Phi Delta Kappan, 83, 758-765.

Stobart, G. (2001). The validity of national curriculum assessment. British Journal of Educational Studies, 49(1), 26-39.

Supovitz, J. A., MacGowan III, A., & Slattery, J. (1997). Assessing agreement: An examination of the interrater reliability of portfolio assessment in rochester, new york. Educational Assessment, 4(3), 237-259.

Testing and Assessment (2008). House of Commons Children, Schools and Families Committee Third Report of Session 2007-08, Volume II. London: The Stationery Office Limited.

Thomas, S., Madaus, G. F., Raczek, A. E., & Smees, R. (1998). Comparing teacher assessment and standard task results in England: the relationship between pupil characteristics and attainment. Assessment in Education, 5(2), 213-246.

Thomas, G., Tagg, A., & Ward, J. Numeracy assessment: How reliable are teachers' judgments? Findings from the New Zealand Numeracy Development Projects 2005.

von Davier, A. A., Holland, P. W., and Thayer, D. T. (2004). The kernel method of equating. New York: Springer.

Webb, N. W. (1993). Collaborative group versus individual assessment in mathematics: Processes and outcomes. Educational Assessment, 1, 131-152.

Wiliam, D., Lee, C., Harrison, C. and Black, P. (2004). Teachers developing assessment for learning: impact on student achievement. Assessment in Education 11(1), 49-65.

Williams, E. J. (1999). Developmental reading assessment reliability study. Retrieved from http://www.pearsoned.com/RESRPTS_FOR_POSTING/READING_RESEARCH_STUDIES/R11.ResearchPaper_DRA.pdf on 10/12/08.

Wilmut, J. (2004) Experiences of Summative Teacher Assessment in the UK. A review conducted for the Qualifications and Curriculum Authority. London: QCA

Wright, D., & Wiese, M. J. (1988). Teacher judgment in student evaluation: A comparison of grading methods. Journal of Educational Research, 82(1), 10-14.

Zimmerman, D.W., Zumbo, B.D. and Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. Educational and Psychological Measurement, 53, 33-49.

QCA Evaluation Reports

1. Monitoring Children's Progress Project (2005). Evaluation Report
2. Qualifications and Curriculum Authority (2006). Monitoring pupils' progress in English at key stage 3. England: Crown copyright.
3. Qualifications and Curriculum Authority (2006). Monitoring pupils' progress Project, Evaluation Report
4. Monitoring Children's Progress Project (2006). Evaluation Report
5. Minnis, M. (2007). Evaluation of the Monitoring Children's Progress Pilot Project 2006-7, Report for Qualifications and Curriculum Authority.
6. Monitoring children's progress pilot project (2007). Evaluation Report
7. Minnis, M. (2007). Evaluation of the Monitoring Children's Progress Pilot Project 2006-7, Report for Qualifications and Curriculum Authority.
8. Minnis, M. (2008). Evaluation of the Assessing Pupils' Progress (APP) in year 1 pilot project, Report for Qualifications and Curriculum Authority.
9. Report on trial of models of moderation within APP (2008).
10. Minnis, M. (2008). Evaluation of the Assessing Pupils' Progress (APP) in key stage 2 Pilot Project 2006-2008, Report for Qualifications and Curriculum Authority.
11. Evaluation report on second pilot project (2008). Assessing Pupils' Performance in Speaking and Listening at key stages 1, 2 and 3
12. Department for children, schools and families (2008). Evaluation of the making good progress pilot Interim report, Research Report DCSF-RR065, United Kingdom: PricewaterhouseCoopers LLP.

Appendix 1: Summaries of APP Pilot Evaluations

Report 1

Monitoring Children's Progress Project
12 December 2005 – 2006
English and Mathematics
Key Stage 2
Number of schools = 42
Number of children per teacher = 4 -6
Number of children overall = 133

Starting the Project

Teachers' were asked to complete a questionnaire after their training day, in general teachers were clear about the aims of the project and were keen to improve their assessment skills. However there were still a large number of teachers who felt they only 'somewhat' understood the aims of the project and felt confident enough to make judgments. Understanding of the aims and requirements was higher for the mathematics group whereas confidence in the ability to collect evidence and make level judgments was higher in the English group. In Dudley LA at least three schools withdrew from participation in the project due to a lack of understanding.

Mathematics

One difficulty noted by the mathematics group was that although one to one discussion with individual pupils was a good source of evidence this could often not be achieved during lessons.
Despite attending training days teachers did not always follow the recommendations in regards to obtaining evidence, and although they were told it was not necessary to accumulate copies of work some did. This may be due to confusion about how to collect evidence and what is classed as evidence. Teachers also found some strands and targets harder to collect evidence for than others and only half made use of the flow diagram provided.
There were some problems with the assessment guidelines including the layout, reviewers also noted that there was some variation in the understanding of the sheets. They were seen by most teachers as being helpful in terms of giving an overall picture of what the child can do and what target setting to make. However there were time and manageability issues, several said it was time consuming. Overall teachers found the project useful and felt they knew more about their target children and had a better understanding of the performance at levels 3 and 4.

English

One problem identified by the teachers was the difficulty obtaining independent work in evidence for writing; this was also seen as problematic in the mathematics group. In reading assessment the major problem was how to find appropriate evidence.
The English group displayed mixed attitudes towards the assessment guidelines; whilst some found them helpful others found they did not help their decision making as they were too complex.
Teachers also doubted the accuracy of their judgments particularly in reading; even by review day 5 out of 20 participants were still unable to identify the differences between level 3 and 4 in reading.
Again there were problems with time issues and manageability, in particular teachers found that to make judgments on reading they had to create specific assessment opportunities. However teachers also said that making judgments became easier and faster over time. Like the mathematics group they also felt that they had learned new things about their children.

Overall

- After initial training some teachers did not fully understand what was required
- They found making judgments became easier however some still found it hard to commit to a level judgment
- Problem with collecting evidence and what is evidence
- Practical requirements need to be made more explicit
- Assessment guidelines could be made more simple as they found it too hard to differentiate between levels
- General enthusiasm
- Don’t feel it can replace other assessment system therefore it is just an additional burden

Report 2

Monitoring pupils’ progress in English at key stage 3
2003 – 2005
English
Key Stage 3
Number of schools = 91
Number of LA’s =15
Number of teachers per school = 2

Number of children per school = 2 x 25

The assessment of ongoing work in MPP

Teachers found the use of AF's positive and worthwhile and it has increased their confidence in making judgments. The time it took to apply guidelines to pupils reading and writing ranged from 2 – 25 minutes with the average being 6 – 11 minutes, teachers found that the time it took to make judgments sped up as their familiarity and confidence in the AF's grew. Teachers believed this system to be superior to current assessment systems. They found the guidelines manageable, not too time consuming and useful.

The use of assessment tasks

Attitudes towards assessment tasks have shifted over the last two years due to improvements in task design and growing familiarity with MPP. Teachers rated the interest of the tasks fairly highly although they were rated slightly lower for accessibility and enabling pupils to show what they could do. Teachers also found tasks easier to assess than ongoing work. On average it took teachers around 4-6 minutes to mark a task. Most teachers found that the tasks were beneficial but only if they were used alone, they also said that they would make use of tasks in future. However tasks are only there to play a supportive role to the ongoing work.

Combining level judgments

Combined level judgments look at combining the ongoing work collection with the assessment tasks to come to an overall judgment. Average times for these in reading and writing were generally 3-4 minutes although some reported times of 2-3 minutes in the last cycle for pupils in year 7 and 8. Ongoing assessment judgments and assessment task judgments were in agreement in around 40% of pupils in reading and 45% in writing. This meant for the others teachers would have to decide which level best represented the pupil. They found that the teachers were more likely to award the level from the ongoing assessment. However over the two years teachers began to take more consideration of the outcome of the assessment task. Sometimes they gave more specific AFs than ongoing assessment did.

Some markers found that when pupils did not reach the expectation of the teachers the teacher deviated in the pupils favour. This results in a need to make the evidence for basing judgments clear to teachers and pupils.

Relationship between MPP level judgments and other national curriculum assessments

In June 2005 the percentage of outcomes from the progress tests and national curriculum levels that were the same were 62% (n=79) in reading and 56% (n=80) in writing. In the cases that were not in agreement the ongoing work in writing gave a higher level than that on the achieved test and vice versa for reading.
In the optional tests for year 7(4, 5, 6) agreement between these results and ongoing teachers assessment was 48% (n=413) for reading and 57% (n=383) for writing. In the results that did not agree teacher assessment levels were higher than those of the optional test.

The optional results for year 8 were also similar with agreement at 52% (n=432) in reading and 57% (n=400) in writing, as with the year 7 results the teacher assessment level was higher in those results that did not agree.

In 2004 year 9 key stage three level judgments were in agreement with ongoing teacher assessment for 61% (n=630) of pupils in reading and 70%(n=639) in writing. Where there was disagreement in writing teacher assessment was likely to be higher than the test level, for reading teacher assessment was just as likely to be higher as it was to be lower.

In year 9 reading 2005 teacher judgment levels and the outcome from the national tests were the same for 56% (n=676) of pupils and for writing 59%(n=640) of pupils. When the outcome was different teacher assessment was more likely to be higher than the test level.

Tendency for teachers' judgment of levels to be higher than those achieved on tests. This may be because they are looking at aspects the tests are not, this highlights the issue of ensuring teacher judgments are made with an understanding of the features of performance at each level and the need to make sure teacher assessment is standardized and moderated.

Reliability and Consistency in MPP

Overall a level judgment made by different judges on ongoing collections of work has been consistently at around 40% for agreement at sub level and 70% agreement for whole levels.

There has been greater agreement in writing than reading, sub-level -writing 58%(n=24), reading 41%(n=35), one whole level - writing 86%, reading 80%. However this small sample must be taken into consideration.

Moderation sessions can help to make assessment more accurate by bringing issues into the open such as making assessments better to encourage weaker pupils, or filling in gaps instead of noting they had insufficient evidence.

Using MPP to track progress through key stage three

In 2003/4 73% (n=532) of pupils in reading maintained or improved their sub level and as did 77% (n=578) of writing pupils. The remaining pupil's pattern of achievement was more erratic. Some pupils actually achieved a lower level in the progress tests than in the key stage 2 tests the previous summer. This is more obvious in reading than writing; it may be because in reading at ks2 pupils achieve a higher level therefore there is less room for improvement, it also may be because teachers spend less time on developing reading.

MPP has the ability to show patterns of progress, teachers welcomed this. However it does not lead to fast improvements as teachers need to come to terms with different ways of doing things. It can tell us about pupils who do not achieve well over time and planning and teaching can help to improve this.

Insights from MPP into teaching and learning in key stage 3 English

Work on reading at KS3 focuses on AF 2 and 3 but not 4,5 or 6, at year 7 pupils perform better on AF 2 and 3. As reading has not previously been distinguished form writing before this came as a change to teachers, but it has highlighted the importance of developing reading .

Writing work was more plentiful; however it was often hard to evidence weaker pupils as it was heavily scaffolded. Teachers had some problems with the AF's as they were not as consistent as those for reading.

Overall

- Discrepancy between levels in task assessment and ongoing assessment, what are they measuring?
- Weakness is that it is hard to determine what led to a particular judgment, evidence needs to be made clear
- Tendency for teachers' judgment of levels to be higher than those achieved on tests.
- Teacher judgments need to be made with an understanding of the features of performance at each level
- Teacher assessment must be standardized and moderated
- Overall a level judgment made by different judges on ongoing collections of work has been consistently at around 40% for agreement at sub level and 70% agreement for whole levels
- There has been greater agreement in writing than reading however the sample is small
- Moderation systems can help support accurate judgments

- Some pupils actually achieved a lower level in the progress tests than in the key stage 2 tests the previous summer particularly in reading
- Teachers admitted paying insufficient attention to teaching and assessing reading
- MPP has the ability to show patterns of progress, it can tell us about those pupils who do not achieve well so teachers can look at new ways planning and teaching can help to improve this, this will, however, take time.
- Work on reading at KS3 focuses on AF 2 and 3 but not 4,5 or 6
- AFs in writing were not consistent

Report 3

Monitoring Children's Progress report
Second evaluation report
2005 – 2006
English and Maths
Key stage 2
Number of schools = 42
Number of teachers = 26
Number of children per teacher = 4 -6
Number of children overall = 133

Spring term findings – Southampton, Birmingham,. Dudley, Croydon, Bromley

Evidence to support teacher assessment

In maths teachers viewed written work and individual work as being most important, in Ma 1 they found written work as being crucial however this was reported as the most difficult to obtain. Teachers had to make some significant changes to their teaching to obtain the evidence required for teacher assessment. In English teachers found it hard to evidence reading and had concerns about the range and quality of the evidence. Some teachers developed questions for use during reading practice which they felt was an extra burden. Others who expended reading experiences had a more positive experience. Teachers felt more confidence finding evidence for writing, however project reviewers are worried about the extent some pieces of writing are scaffolded.

Making level judgments

Although the majority of English teachers said they used the flowchart to help complete their assessments, guideline sheets that were collected were not used as required; they often did not record the AF or strand level. In maths almost half of teachers did not refer to the flowchart or the NC level when making their judgment. A further problem was that teachers were not certain as to whether children would retain new maths skills.

Confidence in level judgments

Although most teachers were generally confident in making judgments about English standards, they did express concerns in regards to reading. These were related to not having sufficient evidence and selecting the best fit level for each AF. In maths teachers were more confident however they still had problems with getting sufficient evidence, selecting the best fit level and making the decision as to whether a pupil was secure, low or high.

Teacher Response to the Project

The majority of teachers feel that the MCP tells them more about their pupils than using a test and that it is worthwhile way of assessing pupils. Teacher's management of the project improved over time and they generally expressed positive statements.
In terms of using MCP the majority said they would not know how to assess a whole class. Reasons for this were that it was time consuming and would be another burden on top of other exams. On the other hand nearly two thirds said that it might fit it to the school process for setting individual and group targets. Also almost two thirds agreed that individual teachers could use MCP in their own classrooms however they felt if it was adopted a whole school model would be best. Nearly all teachers agreed that if MCP was introduced it needs support from senior managers to work and structured training would be necessary.

Sefton and Wirral review of first term

No. of teachers = 30

Teacher Response

The majority of teachers found the MCP approach beneficial to finding more out about their pupils. This group were more open to the MCP approach and more willing to incorporate into current school assessment, this maybe because teachers within these schools were more likely to have discussed the project with other colleagues or had meetings.

Ongoing issues for the project included, lack of evidence for reading and ma 1, limited opportunities for independent work, unfamiliarity with criteria and the time assessment takes.

Making level judgments

All teachers were confident in making level judgments in maths and most were confident in English, however when reviewers analysed some of the judgments s they found that they may have been made using limited evidence.
These teachers worked with revised materials which they found easy to use but had to be reminded by reviewers to refer back to the flow charts and guidance materials.

Assessment outcomes

Mathematics – n=101
Reading – n = 124
Writing – n = 145

In Ma1 nearly a third of pupils were judged as being level 4, just under half were judged as level 3. In Ma2 51% of pupils were most likely to achieve a level 4. Pupils were most likely to achieve a level 4 in Ma 3 and 4. In reading 47% of pupil were judged as level 3 and 51% at level 4. In writing 44% of pupils were awarded a level 3 and 52% a level 4.

Issues for discussion

Teachers beginning the MCP approach tend to have similar problems such as lacking familiarity, lacking confidence, blaming the process, doubting their own expertise and getting hung up on recording. Teachers can become more settled into the process with support.

Teachers consistently feel that they cannot offer all the activities to address all facets of the subjects as described in the strands and AFs. Their time is always being squeezed to produce good test results and to cover new subjects.

In reading teachers admitted that they were often not reading explicitly and there appears to be a constant issue in the lack of independent writing.

Due to the culture of testing teachers have not been able to experience quality teacher assessment, they lack confidence and rely on testing due to the 'high stakes' of assessment, teachers and head teachers need to see the programme as worthwhile to the development of their pupils.

Overall

- In English teachers found it hard to evidence reading and had concerns about the range and quality of the evidence

- Teachers felt more confidence finding evidence for writing, however project reviewers are worried about the extent some pieces of writing are scaffolded
- In maths teachers were more confident however they still had problems with getting sufficient evidence, selecting the best fit level and making the decision as to whether a pupil was secure, low or high
- The majority of teachers feel that the MCP tells them more about their pupils than using a test and that it is a worthwhile way of assessing pupils.
- The majority of teachers said they would not know how to assess a whole class using MCP but it could be used for individuals or groups
- Ongoing issues for the project included, lack of evidence for reading and ma 1, limited opportunities for independent work, unfamiliarity with criteria and the time assessment takes.
- Teachers do not make full use of the flow chart or guidance material
- Culture of testing, results in teachers lacking in confidence and relying on tests due to 'high stakes'.

Report 4

Monitoring Children's Progress Project
End of summer term review meeting
25 July 2006
English and Maths
Key Stage 2
Number of schools = 42
Number of children per teacher = 4 -6
Number of children overall = 133

Teacher Questionnaires – Summer term

Availability of evidence of pupils' skills and understanding has been a problem throughout the project.

All teachers grew in confidence and by the end of the summer term they all said that they were confident in their judgments although more so in maths. There were specific concerns in Ma 1, 1 to processing in Ma 4, 1 to Ma2 and 1 to mental mathematics. In English there were concerns in AF7 in reading, spelling, AF 5 in writing and AF6 in writing.

The number of guidelines returned in this round was less and the % of no judgments was higher, also the quality of the guidelines returned was so low and hurried that they couldn't be sure if they were quality judgments.

Teacher perceptions of MCP

Both teachers from Maths and English most frequently selected the benefit that the MCP gave a greater understanding of the characteristics of attainment at levels 3 and 4. Teachers in the English groups were more likely to choose the ability to identify good assessment opportunities and knowing the evidence to support assessment judgments than those in the maths group. (see paper for stats)

48% of teachers in the maths group said that MCP had helped them to identify gaps in the curriculum compared to 38% of English teachers.

70% of maths teachers and 72% of English teachers said that using detailed assessment of individuals would be an effective or very effective indicator of whole class attainment.

Note: it takes 2.5 weeks of marking time to mark the y5 optional English tests.

Taking MCP further in schools

10 schools in the mathematics group were in discussion as to whether to take MCP further 2 planned to and in the English group 6 planned to. Teachers were asked if they had any plans to continue with parts of the MCP, in the English group 81% said yes and in the maths group 83% said yes. English teachers were more ambitious in their plans.

Assessment Outcomes

More levels 4's are given in M1

Judgments for AF level for reading and writing

There are more level 4's given in AF 4 and 5, there were less in AF 2, 3 and 7. This may mean that teachers need to be more secure in their AFs for reading.

For both maths and English there was a rise in the number of cases where no judgment was made in the spring term, this may be due to teachers not appreciating that the purpose of the review day was to collect completed data.

Level Judgments for Maths

51% of children were judged to have achieved a level 4 in Ma2 compared to 31% in Ma1 (spring) Ma 3 and 4 also performed strongly, Ma 1 improved in the summer term (52%), this may be because of the work done in the review meetings.

Level judgments for English

Performance in reading is poorer than writing. There was also a larger number where no judgment was given this may be due to a loss of focus on completion of the guidelines.

Teacher assessment and optional test results

Percentage of agreement for writing (n=40) was 33% and reading (n=32) was 41%, for both the test result was more likely to be higher than the assessment sub-level, however the sample was small. Maths (n= 49) Ma1 was in agreement with test results at 29% if there was a difference this was due to test result being lower that the teacher judgment. Ma2 = 21%, Ma3 = 25% and Ma4 = 23%, if there were differences here it was more likely that the test was higher than teacher judgment.

Discussion

Issues that need to be considered are; time that it takes for teachers to come to terms with the programme, distinguishing between the projects effects and the MCP effects, limitations in curriculum coverage and finding assessment opportunities and pressures of existing policies.

How can the objectives of the project be achieved without overwhelming the participating teachers?

Teachers have found the change challenging and sometimes even unsettling, some teachers dropped out of the project, those who stayed did not always get things right.

How can schools gain a clear understanding of the project requirements?

Often schools did not fully understand the requirements of the project; teams should tell them exactly what is required in advance, by being ‘upfront’ this should limit confusion although they will probably still need reminders.

How do teachers visualise MCP?

Teachers were unsure how to visualise the MCP working as they saw it more as project work. The team needs to introduce a clearer picture as to what the MCP is and the benefits it can bring.

Are there effective ways to model good teacher assessment to support teachers in the project?

The English teachers found the review meetings useful, they had a chance to be ‘pupils’ and practise the activities.

Suggestions from teachers

Become familiar with guidelines, head teacher support is crucial, importance of planning should be explicit; also record things they say, give teachers practical examples of assessment activities, have a whole school approach

Overall

- Availability of evidence of pupils skills and understanding
- Difficulty obtaining evidence for reading
- The quality of the guidelines returned in this round was so low and hurried that they couldn’t be sure if they were quality judgements
- More levels 4’s are given in M1
- teachers need to be more secure in their AFs for reading
- Performance in reading is poorer than writing.
- Teachers have found the change challenging and sometimes even unsettling
- Need support from head teachers
- Planning is important

Report 5

Evaluation of the Monitoring Children’s Pilot Project 2006-7
February 2007
English and maths
Years 3, 4 and 5
Key stage 2

Number of schools = 104
Number of children per teacher = at least 6

Evaluation findings

Observation of training events

Teachers responded positively to the training sessions, in the first there was still a feeling of being overwhelmed, by the end of the second teachers seemed to understand what was required of them as they had more opportunity to work in pairs and groups to engage with materials.

Analysis of questionnaire responses

119 completed questionnaires were received from 52 schools, 46% coming from schools doing English and maths, 29% doing English and 26% coming from maths.

Completing the assessment guidelines

The average time for completing the guidelines in maths was 53 minutes per pupil and for English was 74 minutes, they ranged from 20 minutes to 4 hours. Teachers said it took long time on this occasion because they needed to seek evidence and were unfamiliar with the process. English teachers found the process harder than maths teachers. English teachers found obtaining evidence the most challenging aspect whereas maths teachers found making a judgment just as challenging as finding evidence.

In reading AF 7 was the hardest to find evidence for, in writing AF1 and 4 were identified as being most problematic. In maths Ma 1 was considered to be the most challenging particularly reasoning. In Ma3 teachers found position and movement most difficult, and in Ma 4 AFs for processing and interpreting were most difficult.

The majority of teachers viewed their judgments as reasonably accurate (not very accurate) for English and maths.

Learning from assessment outcomes

60% of teachers said that carrying out the assessments had added to their knowledge of the children, half said it gave them insights into strengths and weaknesses of the child. The majority also said that the project had improved their understanding of what is expected at NC levels.

The activity teachers found most useful was completing guidelines, this was chosen most by maths teachers. The next was in school moderation which was mostly picked by English teachers.

95% of maths teachers said that the project had given them useful information about their class as a whole, and 75% of English teachers also agreed. Reasons they gave for this included identifying gaps in learning and the curriculum.

The majority of teachers were eager to use this information in subsequent teaching.

Processes in the MCP project

Nearly all teachers were involved in some in-school standardisation and just fewer than 80% were involved in in-school moderation. Only a few teachers involved other colleagues in the process, this reflects the schools limited view of the project.

80% of teachers said they found the completion and submission form clear and easy to follow, although a substantial number of maths teachers said they had not, this may have been because they could record ‘insufficient evidence for an attainment target.

The majority of teachers found moderator visits quite or very valuable, also the majority of teachers considered the head teacher to be supportive although not many were closely involved.

Data collection and the sample

The sample included 83 schools (54 English, 56 maths), 1342 pupils. This represents the MCP study as a whole.

The assessment guidelines are based on a decision between levels 3 and 4, below 3 refers to level 2.

English year 3

The strongest AFs in reading are 1, 2 and 3. Insufficient evidence was a significant problem in AFs 5, 6 and 7, particularly 7. In writing insufficient evidence was not a problem, but there were problems with AF4 and AF6.

English year 4

Again the strongest AFs in reading are 1, 2 and 3. Insufficient evidence was a significant problem in AFs 4, 5, 6 and 7, particularly 7 where 42% of pupils had insufficient evidence. In writing the strongest AF was 8 and lack of evidence was significant in AF 4 and 6.

English year 5

The same applies in reading as in year 3 and 4, again AF2 and 3 are the strongest. In writing the weakest AFs are 4 and 6, performance in AF 2 is stronger than the previous year.

Maths year 3

In Ma1 there was insufficient evidence for at least 20% in all AFs, communication was the strongest. In Ma2 the best performance was in numbers and the number system, the weakest AFs were operations, solving numerical problems and written methods. In Ma 3 performance was strongest in properties of shapes, in position and movement insufficient evidence was recorded for more than half the pupils. In Ma4 over 30% recorded insufficient evidence for all 3 AFs.

Maths year 4

Again the strongest AF was communicating in Ma1, in the other 2 AFs there was insufficient evidence for around 30% of pupils. Again in Ma2 best performance was in numbers and the number system followed by mental methods, the weakest AFs were operations, solving numerical problems and written methods. In Ma3 performance was weakest in position of shapes and movements although this may be because there was insufficient evidence for over half the pupils.

Performance was strongest in properties of shape. In Ma4 all insufficient evidence was recorded for just under 30% of pupils in all 3 AFs.

Maths year 5

Again the strongest AF was communicating in Ma1, in the other 2 AFs there was insufficient evidence was slightly better. Again in Ma2 best performance was in numbers and the number system, there was a significant lack of evidence for operations. Written methods are stronger in year 5 and there is less difficulty finding evidence. In Ma3 there were 66% of pupils lacking evidence for position and movement. In Ma4 there was a higher recording of insufficient evidence for representing AF.

Overview of the summary reports from moderators

Mathematics

Moderators thought that teachers were making consistent and mostly accurate judgments, however they were sometimes based on confident expectations as opposed to the child's actual ability.

In Ma1 moderators were in 87% agreement with teachers, in Ma2 agreement was 83%, the most disagreement was in fractions.

Evidence was lacking in all AF's in Ma1 and in mental methods, solving numerical problems and written methods for Ma2. There was little evidence of independent work and within schools the range of evidence was limited.

In-school moderation worked well, however moderators noted that there was lots of 'discussion' but not many differences in judgment.

English

Teachers were generally accurate and consistent more so in writing than reading. Where in accuracy occurred it tended to be due to lack of understanding, uncertainty and inadequate evidence.

In reading moderator and school agreement ranged from 59% (AF3) to 95% (AF1). In writing moderator and school agreement ranged from 74% (AF3, 6) to 86% (AF8).

In reading there was a significant lack of evidence. Moderators also felt that the independent work was too structured.

Although teachers found in school moderation helpful in terms of understanding the assessments moderators found that there was little challenge to judgments made within schools, instead teacher sought reassurance in their judgments.

Discussion

Teachers were enthusiastic and their confidence grew throughout the process despite the project being challenging. They often found the work too much as collecting evidence takes a lot of time and effort. However most teachers implied the outcome was worth the effort.

Although teachers said they felt accurate in their judgments moderators felt that lack of evidence in both quality and quantity was a serious issue, 60% of teachers also said that collecting evidence was the most difficult aspect.

Moderators were also concerned that teachers avoided the trickier aspects of the assessment process.

Emerging issues

There were difficulties with teachers at the beginning of the project in understanding the nature of assessment, teachers need to understand the concept of the purpose and understand clearly from the start.

Teachers who are keen to get assessments right will devise ways to do this, e.g. Create questions about specific AFs. There is confusion about how to integrate assessment into classroom practices e.g. some dedicate lessons to one AF. Although teachers have found using a small group helpful in understanding whole classroom ability, teachers would have to produce detailed judgments on all pupils in their classroom as expected by parents and senior managers.

Overall

- It took some teachers a long time to complete guidelines on this occasion because they needed to seek evidence and were unfamiliar with the process.
- English teachers found the process harder than maths teachers
- In reading AF 7 was the hardest to find evidence for, in writing AF1 and 4 were identified as being most problematic.
- In maths Ma 1 was considered to be the most challenging particularly reasoning
- The majority of teachers viewed their judgments as reasonably accurate (not very accurate)
- Over half of the teachers said the project had given them new insights
- The majority said the project had given them useful information about their class as a whole and would use the information they found out in the future
- However using a target group may not be desirable outside the project
- The majority of teachers considered the head teacher to be supportive although not many were closely involved
- Insufficient evidence is a serious problem in the mathematics assessment
- Sometimes judgments were based on confident expectations as opposed to the child's actual ability

- Agreement between moderators and teachers was high for both subjects
- In school moderation for both subjects was helpful for teachers but moderators felt it did not challenge their judgments
- Teachers said they felt accurate in their judgments; moderators felt that lack of evidence in both quality and quantity was a serious issue
- The purpose of the project needs to be clear from the start
- There is confusion about how to integrate assessment into classroom practices

Report 6

Monitoring children's progress pilot project
February-May 2007
English and Maths
Years 3, 4 and 5
Key stage 2
Number of schools = 104
Number of local authorities = 12

<u>Analysis of teacher and moderator judgments</u>

A number of substantial improvements in teacher judgments since report 1. Ma1- AF's = minimum of 94% agreement, 93% agreement on overall grade. Ma3- AFs = min 92%; overall grade 89%. Ma4- AFs = min 93%, overall grade 85%. Reading – AFs = min 85%; overall grade 75%. Writing- AFs = min 81%; overall grade 73%. Levels of agreement for all AFs for reading and writing have improved. Lower agreement for overall level may be due to moderators claiming a lack of evidence to support some judgments/narrow focus on specific AFs therefore claimed not appropriate to award an overall level.

<u>Assessment processes</u>

Using evidence from other curriculum areas can add to teacher understanding of what pupils can do, around 60% of teachers amended their teaching plans to identify where there were opportunities to assess across the full range of AFs.

The majority of teachers found the moderation process either as useful as or more useful than last time, in-school moderation occurred in the majority of schools this time round.

Positive findings

83% of teachers reported using a wider range of assessment activities

General increase in the number assessments at level 3 or above over the two month period.

Teachers found implementation of the MCP very challenging at first, but most found it much easier for round 2. Time taken to assess each pupil/subject fell by 30% for maths and English over the period although there was a wide range of time taken to assess amongst teachers ranging from 10 minutes to 3hours.

Pre-planning for the assessment had allowed the improvements in the quality of the assessments.

Teachers in English have been given more flexibility in deciding on the focus for each moderation round.

Head teachers are hoping to reduce the extent of regular testing as expertise in teacher assessment increases.

Problems encountered

A potential risk that the assessment focuses will be treated more like a checklist or teaching objectives, 30% of respondents were unclear of the difference between a teaching objective and an assessment focus.

Few teachers had extended the MPC processes being used to a wider group of pupils

Moderation: teachers expressed concern over the differences in expectations between different moderators and the difference between maths and English moderation models. Some moderators, particularly English, expect schools to provide evidence that they can access without the teacher present.

Overall

- A number of substantial improvements in teacher judgments since report 1
- Teachers found the moderation process useful
- Teachers found implementation of the MCP challenging at first but this improved over time
- Planning assessment improves the quality

- teachers expressed concern over the differences in expectations between different moderators and the difference between maths and English moderation models

Report 7

Evaluation of the Monitoring Children's Progress Pilot Project 2006-7
September 2007
English and Mathematics
Key Stage 2
Years 3, 4 and 5
Number of schools = 104

Evaluation findings – realisation of benefits

93% of head teachers said that MCP had improved the quality of teacher assessment in terms of progression, greater accuracy and teacher confidence.

Accuracy

Teachers reported increasing confidence in their accuracy, in the summer round 99% were confident or very confident in their judgment compared to 92% in the previous round.

In Ma1 overall teachers and moderators agreed on 90% of levels, in Ma2 they agreed on 98% of judgment, they agreed on 93% for Ma3 judgments and 92% for Ma4.

For reading the overall agreement between teacher and moderator judgments was 79% and 69% for writing.

Agreement has generally been better for individual AF's as opposed to whole attainment targets, this may be because teachers do not use the guidance materials as recommended when making a decision on an attainment target. Moderators often had to remind teachers to use the standard files to help support judgments.

Teachers also found the summative judgment to be less important than the AF's as these have an immediate value.

Evidence

In this round there was less evidence that had been heavily scaffolded. However there were still few instances of a choice of purpose being offered to children although there were more examples of work where children had been given more freedom.

Again evidence to support reading has been problematic; however in the summer round there was more evidence of open-ended activities around text and improvement to questioning styles. Demonstrating an increasing recognition that test type question may not be ideal when assessing higher order reading skills.

In maths there was also improvement in the chances the children got to work independently and make their own decisions; however this largely depended on the school. Teachers still need clarification about what counts as independent work and how they can offer more choice.

Impact on planning and teaching

In response to the questionnaire in the summer round 70% of teachers said they changed their teaching round after the previous round of assessment. The most popular changes were to home in on areas of weakness, cover neglected areas, include a wider range of activities and include more opportunities for choice. Overall over 80% of teachers felt there had been changes to their teaching they felt that using MCP had meant their teaching was more focussed and informed and they planned lessons better.

Perceived impact on learning

60% of teachers said that being involved in the MCP had resulted in changes in what their pupils could do. There was a variety of responses as to what these changes were.

Just of 40% of head teachers felt there was evidence of their pupils making better progress. This might be expected as schools had only been working with the MCP for less than a year. Those who felt there had been improvements found it hard to specify what they were.

Integrating MCP into classroom practice

There has been a small decrease in time it takes to complete the assessment for each pupil. The average time for English was 43 minutes and for maths it was 36 minutes. The minimum was 10 minutes and the maximum was 2 hours. Teachers have found finding time difficult, not only for acquiring evidence within the classroom but also for carrying out reviews, familiarising themselves with materials and making meetings for in-school moderation.

Moderators found that particularly during the summer term many schools did not hold in-school moderations. Some head teachers made time for this however many teachers had to do MCP work in their own time. Also teachers spent a lot of time producing work samples, although project leaders said they informed teachers that polished portfolios were not necessary.

English teachers spent on average 79 minutes preparing for external moderation; maths teachers spent on average 61 minutes. They ranged from 10 minutes to 3 hours. This significant time effort has been judged worth it in terms of how much teachers have learnt. The amount of time put into the assessment must not be underestimated when considering how to make sustainable models.

In terms of recording evidence there were a large number of teachers photocopying work, which is not ideal. Teachers need to understand that the purpose of MCP is to gather evidence form everyday classroom activities. One effective way of organising evidence is to annotate assessment guideline sheets.

The teachers in the pilot focussed on a small group of pupils, some found this useful in understanding their class as a whole, others thought there were problems with this model. Teachers were concerned that using the MCP on a whole class would be unworkable but that using a small group would not be 'allowed' due to being inequitable. If teachers considered having a target group they suggested a mixed ability of boys and girls would be preferable.

Integrating MCP into whole school assessment practice

The majority of head teachers said that MCP could fit into their assessment practice. Those that said it would not fit, often already used a termly teacher assessment of some kind. Head teachers thought that the best time of year to carry out the MCP would be termly.

Head teachers existing assessment arrangement are used mostly for teacher performance management, setting targets at school level, making targets for individual pupils and identifying current priorities. 76% of head teachers said that it would be possible for MCP to replace current assessment practices; the most frequent reason for this was that MCP gives more reliable assessment across a range and has stronger links to curriculum planning.

Evaluation of findings - attainment and progress

At assessment focus levels the majority of pupils stayed at the same level, between a quarter and a third went up a level between spring and summer, there was also a proportion who were judged at a lower level.

In terms of sub-levels given by teachers generally between 50% and 73% of pupils improved by at least one sub-level between two assessment rounds, this varies according to year and subject. It is important to remember that as teachers become more familiar with the assessment model the accuracy of their judgments changes and improves.

Moderators commented that the area that needed most improvement was converting AF levels into overall judgment levels for attainment targets and subjects.

Relationship between individual assessment focuses ad MCP assessment outcomes (see paper for stats)

In reading AF3 is the strongest predictor of the overall level judgment in reading at level 3 and there is also a strong relationship between AF4 and level 4 this was not as expected as it is not emphasised in the guidance materials. In writing AF 1, 2, 3 and 5 are consistently strong indicators in overall writing performance.

In Ma1problem solving and communicating both have a strong relationship with the overall level. In Ma2 there is no AF that stands out as having a particularly strong relationship with the overall level. In Ma3 properties of shape is the strongest predictor and in Ma4 as in Ma2 there is no AF that is particularly stronger than the others.

Overall there is no suggestion that over focus on a narrow range of skills results in better performance. The pupils who have done well will also have done well in the AF most strongly related to that level. Depth across all AFs will contribute to overall outcome.

Comparative analysis of MCP level outcomes and optional test level outcomes

In reading the multi level model shows that pupils are more likely to attain a level 3 from the MCP assessment rather than the optional tests, although pupils in year 3 and 4 were more likely to attain a level 4 on the optional tests than the MCP assessment. Overall there was no significant difference in the likelihood of achieving a level 2 result. In all years pupils were less likely to achieve a level 3 in the optional tests; there was no difference in level 4. In year 5 achievements of level 5 was more likely in tests than in MCP.

In maths there was good agreement between the outcomes of the optional tests and MCP assessment at level 4, in the lower levels pupils would tend to achieve higher levels in the optional tests than the MCP assessment.

In reading there were disparities at all levels and in writing at levels 3 and 5. This could either be due to MCP underestimating pupil's ability or optional tests overestimating ability. This test does not tell us which form of assessment is more 'accurate'.

Advantages of periodic teacher assessment as given by the teachers include; giving a more holistic picture of what the pupil can do and giving a more accurate form of assessment. The main drawback was the time required. The major advantages of using tests were ease of use and recognised and accepted assessment of work. Teachers were most concerned with using tests as assessment because they may not be representative of the child's 'true' achievement.

Relationship between individual assessment focuses and optional test level outcomes

The relationship between all the AF's and the optional test outcome is weaker than the relationship between MCP level outcomes and optional tests.

In reading AFs have little relationship to optional test level outcome. This may be because the tests have little opportunity to assess higher order reading skills. In writing the relationship between AF's and test outcome was slightly stronger, the best predictor being AF 5. In maths the strongest relationship was between Ma2 and overall test outcome.

In general MCP and test use give two different perspectives on child achievement. Teachers can use both to develop their understanding about a child so long as they know the different contribution each one makes.

Evaluation findings – what makes MCP work?

Many teachers struggled to find adequate time to fulfil all requirements of the MCP. In the long term schools will need to make systems their own, however it was clear from this study that meetings and activities with external moderators gave teachers incentive to finish the programme. Therefore having an external structure is obviously important.

Most teachers have grown in their confidence of making accurate judgments over the year. The most effective way of conducting in-school moderation has been where another member of staff not responsible for MCP judgments has been involved.

Schools found preparing for external moderation burdensome. In the summer term moderators were asked to emphasise to teachers that polished and overly annotated samples of pupil work were not necessary.

Those teachers who had supportive head teachers tend to have made most progress using MCP and the head teachers have been impressed with the results. Head teacher sponsorship is critical in promoting teacher assessment in schools and for making changes within schools.

Support from LA's varied; it seems that LA involvement and practical support is important if schools are to continue with the project.

Discussion

Teachers make accurate judgments at assessment focus levels however their judgment for attainment levels overall are not so accurate.

In order for head teachers and teachers to gain maximum benefit from the MCP assessment they must use it to track and monitor systems so they can see how pupils are progressing at class, year and the school levels. It also helps to identify problems with curriculum coverage.

MCP is meant to be a periodic review of evidence. It needs to be distinguished from jotting down evidence, or else there is a risk of teachers ticking off a list of what the child can do with no reflection of what this means to their development.

Overall

- Teachers did not make full use of the materials given to help make judgments, moderators often had to remind teachers to use materials
- More independent work and choice to pupils however teachers still need clarification about how to interpret these activities
- The majority of teachers felt involvement in the MCP had led to improvements in their teaching
- Teachers still found acquiring the time to do work for the MCP very difficult.
- Teachers need to fully understand the purpose of MCP in order to record the correct evidence in the right way
- Teachers were concerned that using the MCP for a whole class would be unworkable but using a small group would not be equitable
- Just over three quarters of head teachers said it would be possible for MCP to replace their existing assessment system
- The majority of pupils made progress throughout the year
- Improvement is needed when converting AF levels into overall judgment levels for attainment targets and subjects
- In depth learning across the range of AFs will lead to a better overall level
- There are disparities between optional tests and MCP assessment in all subjects
- Teachers can use both to develop their understanding about a child so long as they know the different contribution each one makes
- Head teacher sponsorship is critical in promoting teacher assessment in schools
- LA involvement and support is important if schools are to continue with the project
- MCP assessment must be used to track and monitor systems progress at all levels to gain maximum benefit
- MCP is meant to be a periodic review of evidence it needs to be distinguished from jotting down evidence

Report 8

Evaluation of the Assessing Pupils' Progress (APP) in year 1 pilot project
September 2007- July 2008
English and Maths
Year 1
Key stage 1
Number of schools = 51
Number of local authorities = 8

Positive outcomes

90% of the teachers (February) considered that APP was giving them useful info about their pupils' learning, 84% APP had improved their ability to identify gaps in pupils' learning. Ability to identify good assessment opportunities and improvements in understanding of characteristics of each NC level were also reported by a majority.
Kinds of evidence used by teachers had changed as a result of APP and a reduction in the manageability of collecting evidence showed an increase.
Many teachers referred to the idea that assessment practice had now been embedded in the planning of teaching.
Use of APP had improved the knowledge that underpins sound assessment practice.
Pupils were experiencing a wider range of learning opportunities with more opportunity for independent learning.

Challenges

Managing the process itself less of an issue than apparent for teachers in the KS2 pilot, although time taken for assessment per pupil did not decrease over the pilot, the average in the summer term was reported at 34 minutes.
Support for year 1 teachers varied widely especially in the time available to work with TAs or the time to train TAs in the AFs.
No significant difference in the manageability of assessing English or maths unlike KS2.
Important that the whole school is signed up to using APP to reduce workload of having to run two assessment processes simultaneously.
Preparing for moderation was seen by head teachers as being the least manageable aspect of the process.
A single day of training to develop an understanding of all of the underlying aspects of APP was seen as less than adequate, these skills only developed across the moderation meetings.
Initial concern in the transfer from assessment schemes laid out as a tick-list of criteria to APP where teachers are required to apply their own professional judgment.

Year 1 specific challenges

Year 1 is the first time that pupils are working within the national curriculum, some pupils enter year 1 ready for national curriculum assessment, others are slower and are better suited to the previously used FSP (foundation stage profiling) and others are in between. As yet there has not been a format provided that allows an easy comparison of the two formats to ease classification of those that are 'borderline'.

Evaluation findings

- Majority reported that the use of structured and consistent assessment criteria within APP had improved the quality of teacher assessment judgments; however these improvements were not instant.
- The use of shared language between KS1 and 2 has increased the trust from teachers in one school to the next.
- 87% of head teachers surveyed believed that APP could replace some or all of their existing assessments in year 1 and others saw it as a means of reinforcing the other assessment practice.
- Some expressed concern regarding the pressure from external agencies to express sub-levels/count points and how this would fit with APP.
- The ease with which a school and teachers can adopt APP will depend upon how well it fits with existing practice and the extent to which the principles for assessment for learning are in place.

Recommendation

Moderation is a vital part of the assessment system. However there should not be too much emphasis too soon as it may lead to confusion over key concepts of APP. Feedback suggests that APP is often best introduced in 'manageable chunks' and built up in stages.

Overall

- APP gives useful information about pupils' learning and improves ability to identify gaps in pupils' learning
- Teachers assessment judgment improved over time
- Time management is still an issue
- Majority of head teachers believed that APP could replace some or all of their existing assessments in year 1
- Important that the whole school is signed up to using APP to reduce workload of having to run two assessment processes simultaneously
- There is pressure from external agencies to produce levels, how would this fit in with APP
- Moderation is vital

Report 9

Report on trial of models of moderation within APP
September 2007 – July 2008
English and Maths
Key stage 2 and 3
Number of collections moderated = 628
Number of local authorities = 14

Moderation background

Moderation of APP takes place in a tiered structure, base-teachers, in-school moderation external moderation (LA level), national review (accuracy check).
The majority of teachers estimated approx. 1.5 hours of preparation time for a moderation session, although this was regarded as 'fairly manageable' it did not decrease across the pilot.
Teachers valued the opportunity to work with teachers from other schools and to work with colleagues in group discussions.

Types of moderation

Postal moderation deemed to have limited use as did not support professional development; small trial of web-based moderation was received enthusiastically.
Regardless of the method of moderation used, each teacher received a brief verbal report and a subsequent written report from a moderator.

Moderation findings

75% and 76% teacher judgment accuracy reported for maths and English respectively.
Collections of evidence for moderation had to be developed from those which supported ones own personal and professional judgment to evidence which could be accessible to others. Ratings of accessibility improved for both English and maths
Valid assessments require the student to work with some independence

Consistency of judgments

More judgments were confirmed as correct in the later rounds of moderation which suggests increasing familiarity and confidence in assessment focuses by the teachers.
Where the moderation judgment did not meet that of the original teacher judgment this was often as a result of a lack of sufficient evidence from the school.

Some AFs were more difficult to provide evidence for than others and thus were less likely to be confirmed on external moderation.
Data reflects that the different rounds of moderation had different assessment focuses from one round to the next
Moderation valued highly by teachers as a source of professional development
Significant improvements found in the quality and range of assessment collections provided by the summer round.

Recommendations

When choosing the moderation model to be used, LAs take into consideration the strengths and weaknesses highlighted by the report of the pilot.
Secondary reviews of selecting samples of moderated collections of evidence can be used to monitor the effectiveness of moderation within an LA, between LAs and also nationally. Systematic review can reveal discrepancies in understanding and practice between moderators.

Overall

- Moderation preparation did not decrease throughout the pilot
- Around three quarters of judgments were accurate
- Judgments improved as teachers became more confident and familiar with the assessment
- Teachers found moderation useful
- Teachers had improved in their range of assessment collections by the summer round

Report 10

Evaluation of the Assessing Pupils' Progress (APP) in key stage 2 Pilot Project
Autumn 2006 – July 2008
English and maths
Years 3, 4, 5 and 6
Key stage 2
Number of schools = 74
Number of local authorities = 12

Year 2 of the pilot

The second year shared the same pilot requirements as previously. LAs had a far more prominent role this time with any teacher training being implemented by the LAs rather than by QCA. Schools were given a wider assessment period window to allow them to carry out the assessments at times more convenient to them.
Changes to APP materials were made over the summer 2007 in response to

feedback from teachers in the first year. Materials were made available to all schools through a national development programme.

Alongside the main KS2 project ran a related project exploring the use of different moderation techniques, most schools chose the form of joint moderation with local schools and colleagues.

The organisational changes made in the second year of the pilot meant that teacher experience of APP was perhaps more similar to what might happen as APP is adopted on a voluntary basis by local authorities or groups of schools. There was more variation in approach to suit local needs, access to local resources and support to supplement the nationally published materials and guidance. This made drawing conclusions about the embedding of APP more difficult.

Positive findings

92% of teachers again reported that APP had improved their understanding of NC levels.

English and maths moderators reported improvements in quality and range of evidence used with reference to the use of evidence from other subject areas.

A vast majority of schools reported how well APP supported not only teaching but also the planning involved in both assessment and teaching.

Fewer ‘anomalous’ grade changes between rounds in the second year of the pilot.

97% of teachers said that APP had enabled them to identify gaps in their pupils’ learning.

Around 60% of teachers felt that they could already see an impact on pupil attainment after using APP for 3 terms.

Less willingness to change classroom teaching practice where APP is seen only as an adjunct to test based assessment.

By the end of the first year of the project, 65% of teachers responding to questionnaires were confident in the accuracy of the judgments they were making using APP and 34% were ‘very confident’.

Using APP guidelines became easier and less time consuming with each round of assessment.

Assessment priorities have changed in a number of the schools in the second year of the pilot so that APP is the main assessment programme with the optional national curriculum tests either being used to confirm teacher judgments or no longer being used at all.

The use of APP has allowed easier communication of a pupil’s learning with parent/carers, there has been an increase in the number of schools using the information for this purpose however this is still not a majority.

Problems

The AFs which were found to be the most problematic were reported to be so because of difficulties in locating good quality evidence; however these problems correlated with areas highlighted by moderators as showing a lack of understanding.

The second year of the pilot generally involved fewer moderation events. Having insufficient time was reported as the most challenging aspect of APP.

Having a highly detailed knowledge of a few pupils and how this can help others in the class at an individual level (rather than in terms of general messages about curriculum coverage etc) remained an unresolved problem for many teachers throughout the pilot.

Overall

- The majority of teachers had improved heir understanding of NC levels
- There was an improvement in quality and range of evidence
- A vast majority of schools reported how well APP supported both assessment and teaching
- APP gives useful info about their pupils' learning and improves ability to identify gaps in pupils' learning
- Over half of teachers felt that they could already see an impact on pupil attainment after using APP for 3 terms
- Most teachers were confident in the accuracy of the judgements
- Using APP guidelines became easier and less time consuming with each round of assessment
- There were still problems with locating good quality evidence; these problems correlated with a lack of understanding
- Having insufficient time was reported as the most challenging aspect of APP.
- Teachers were not clear how having a highly detailed knowledge of a few pupils can help others in the class at an individual

Report 11

Assessing Pupils' Performance in Speaking and Listening at key stages 1, 2 and 3
Evaluation report on second pilot project 2007 – 2008
Speaking and Listening
Key stages 1, 2 and 3
Number of schools = 86
Number of pupils per teacher = 4 - 6
Number of LA'S = 12

Introduction

This pilot is built on an earlier two phase pilot in 2006 – 2007. Initially teachers were given 2 weeks to plan and think about the APP materials. Subject knowledge of speaking and listening although valued at all levels is not understood in terms of progression, which makes a case for the introduction of APP. However it may also involve re-introducing teachers to the curriculum.

Teacher's confidence in planning varied some planning was more ad hoc, others was more formal and some went straight to trying things in the classroom.

Implications for dissemination

Many teachers will need further help with planning, for example a chart to link AFs with curriculum areas otherwise teachers may plan from the assessment focuses. Guidance needs to be clear and to position the APP materials in relation to teaching objectives, in order to present APP as part of a whole picture. Teachers also need to be supported in planning for progression.

Teaching and Learning

The majority of teachers agreed that focussing on speaking and listening has many benefits for children. More recently speaking and listening has been built into teaching, those that were more familiar with the QCA/Strategy materials covering KS 1 and 2 routinely planned for teaching and recording speaking and listening skills. These teachers found additional APP materials useful. However not all teachers will have the knowledge or confidence to design activities for speaking and listening, how to design 'good quality activities' is a concern'.

Some experiences of teaching and learning

Most teachers made some changes to their teaching in order to encourage speaking and listening, although the quality of teaching and learning in speaking and listening has been variable. Most noted that they needed to be doing more focussed teaching to produce the appropriate evidence although they did not necessarily know how to go about doing this.

Implications for dissemination

Teachers need to understand the principles that should underpin god quality teaching of speaking and listening in order to help them plan progressively.

Cross-curricular contexts

At Key stages 1 and 2 teachers used many different lessons to provide opportunities for speaking and listening. At Key stage 3 this was more difficult, although in one school an experiment demonstrated that other subjects at KS3 could also be trained to use the materials and make assessments. At KS3 level there needs to be strong senior management involvement to organise systematic involvement of other subject teachers

Implications for dissemination

There is ample evidence that speaking and listening evidence can be gathered via other subjects at KS 1 and 2, at KS other evidence would be valuable if suitable arrangements could be made. It may be more preferable to promote the APP materials to teachers of all subjects so they could apply their increased understanding of speaking and listening whilst teaching their own subject.

Evidencing the assessment focuses

The majority of teachers said they were able to locate sufficient evidence for their target pupils to conduct assessments for at least the 2 AF's they were focussing on. Evidence differed depending on planning and teaching. Generally teachers were able to observe evidence related to their AF's without difficulty although they did not attempt all 6.

Implications for wider dissemination

There was a problem with AF6 as teachers have often ignored the area of the curriculum relating to this. Also several teachers asked why there are 6 AF's not 4 to reflect the 4 strands in the curriculum. The report suggests that only the strongest schools take up the 6 AF model.

Collecting and recording evidence

The approaches were broadly split into 2: the first captured evidence other than assessment guidelines judgments were drawn later to support the guidelines, the second way was to use the assessment guidelines as the tool for record keeping usually supported by comments or annotations. There were also a variety of methods to go alongside these approaches. Over the pilot teachers improved ways in which they recorded evidence and the confidence in which they recorded.

Three record sheets were supplied by QCA and some made their own record sheets. A few teachers made their assessments onto one of the record sheets provided and some used a combination of both.

Content of teachers' observation notes and comments

Some findings from the pilot included: comments were hard to link up with assessment guidelines, some general descriptions were not indicative of a particular level, some were not relevant and others were largely negative.

Teachers often found it hard to relate what a pupil said and did to the language used in assessment guidelines particularly at the beginning of the pilot.

Overall general description was more common than specific comments of what a child said.

There was a mixed response as to whether video and audio recording were successful. Those who did use it found it very useful to clarify assessment decisions.

Frequency of record keeping

The training materials did not give any guidance as to how often teachers should record pupil work or how often they should assess it. Therefore responses varied considerably from conveying a clear picture to being general and vague.

Key stage 1

They were more likely to use a mixture of recording. The most common frequency of recording was about once every two weeks, a smaller number said once each half term and a few reported between 4 and 8 times over the project.

Key Stage 2

They recorded with less frequency than KS1 and it was more varied, many recorded weekly, fortnightly monthly or half-termly. They were also more likely to report that they had not been able to do as much observation and recording as they had hoped.

Key Stage 3

Again frequency was less than the previous Key Stages; three times across the pilot was a common response. The recording tended to be linked to specific speaking and listening tasks.

How many pupils should records be kept for?

Feedback from teachers suggests that it would not be sensible to keep a record sheet system for every child. What they learnt from their target group may have helped their awareness of the capabilities of the rest of the class.

Assessing speaking and listening: suitability of materials to support assessment

It took teachers some time to become familiar with guidelines. Many felt more secure about their judgments after a term and a half of using the guidelines.

A minority thought the guidelines were too complicated and should be simplified; overall most thought that after enough time and practice they could use the guidelines to make judgments.

Teachers felt that generally they did not need huge amounts of evidence to make a secure level judgment, to do this they need an understanding of the AF's and the assessment guidelines.

Implications for dissemination

This pilot has shown that to make meaningful assessments teachers need to understand the assessment guidelines for each AF and this can be difficult for many. There is a need for training to make sure teachers are clear in their understanding of assessment guidelines, particularly as there is no established pedagogy for speaking and listening. However once teachers had become familiar with the guidelines they found them useful in telling them things they did not already know about the pupils.

Assessing speaking and listening: manageability of process

Teachers suggested that to be confident about their judgments evidence must: be of a high quality related to planning and showing what the child can do, be based on more than one occasion, be given enough opportunity for the child to produce, be produced between three and five times per child, 'formal' evidence is essential, drawn from across the curriculum and teachers must use their professional judgement in order to be secure.

Just over 40% of teachers thought that assessing pupils was manageable, just under 50% conducted fewer assessments than they had planned because of practical difficulties and 10% found it hard to conduct assessment throughout the whole pilot.

Making assessments for a whole class

Less than a third of teachers felt they fully understood all assessment focuses and guidelines and could assess all pupils in their class so long as it was only twice a year. Under half thought that assessing a whole class would be manageable as long as they weren't assessed at the same time and no more than once a year.

Moderation

The pilot did not include trialling any moderation procedures although teachers were encouraged to share evidence and judgments, from a few limited example of moderation it was found that teachers could explain judgments and why they had made them. There were no significant disagreements suggesting that teachers had learnt to apply the guidelines in the same way. However it wasn't possible to tell if a whole group would have agreed on the same thing.

Teachers views on what would be needed to support the wider use of APP

materials

Suggestions from the support for the planning and teaching of speaking and listening includes: having exemplar plans and activities, having banks of teaching ideas, resources to support teaching, training for teachers in teaching speaking and listening before attempting to assess it.

Guidance and training to support assessment

Suggestions include: training for teaching assistants, advice on manageable record keeping, guidance on whole class assessment, guidance on standardisation and moderation.

The majority of suggestions were to do with the practicalities of how to teach speaking and listening and how to assess it in manageable ways.

Overall

- Teachers may still need guidance with planning
- Guidance must be clear and present the APP as part of a bigger picture involving planning, teaching, assessment and review
- How to design ‘good quality activities’ is a concern’
- Most noted that they needed to be doing more focussed teaching to produce the appropriate evidence although they did not necessarily know how to go about doing this
- The majority of teachers said they were able to locate sufficient evidence for their target pupils to conduct assessments
- Teachers asked why there are 6 AF’s not 4 to reflect the 4 strands in the curriculum
- Over the pilot teachers improved ways in which they recorded evidence and the confidence in which they recorded
- The higher the key stage the less recordings were made
- Feedback from teachers suggests that it would not be sensible to keep a record sheet system for every child
- A minority thought the guidelines were too complicated and should be simplified
- Generally teachers were positive
- Teachers felt that generally they did not need huge amounts of evidence to make a secure level judgement

- There is a need for training to make sure teachers are clear in their understanding of assessment guidelines
- Nearly half of teachers couldn't conduct all the assessments they had planned due to practical difficulties
- Under half of teachers thought that whole class assessment would be manageable and if so, only once a year
- There were no significant disagreements amongst judgements suggesting that teachers had learnt to apply the guidelines in the same way
- Teachers need teaching speaking and listening before attempting to assess it
- The majority of suggestions were to do with the practicalities of how to teach speaking and listening and how to assess it in manageable ways.

Report 12

Evaluation of the making good progress pilot Interim report
June-July 2008
Key stage 2 and 3
Number of schools = 450
Number of head teachers (interviewed) = 26
Number of head teachers (surveyed) = 115
Number of children (interviewed) = 430
Number of parents/carers (surveyed) = 628

Project logistics

All MGP schools divided into three samples: 10 'deep-dive' schools; 40 'light-touch' schools; and the remaining population of pilot schools. Deep-dive methodology involved: interviews with key staff; pupil survey/focus groups; short teacher survey; survey of a selection of parent/carers. Light-touch involved: telephone interviews with head teacher and/or School pilot leader; and a survey of a selected 100 parent/carers. Head teachers of the remaining pilot schools received an online e-survey. In addition all LA pilot leaders were interviewed.

Problems encountered

The majority of schools share APP criteria with their students in some way i.e. pupil friendly versions, although concerns have been raised over potential inaccuracy of these 'translated' versions. However, it is reported that in most cases APP criteria had not been shared with parents and carers.

Three major challenges highlighted relating to APP criteria: increased teacher workload; embedding of APP criteria for formative assessment; and inconsistent use of APP criteria within and across schools. More LA support has been suggested. It was reported by a number of teachers that short-term increases in workload have occurred but in the long-term it should decrease their workload.

Still a tendency to rely on traditional summative assessment tools other than the APP criteria, particularly in secondary school mathematics.
Positive findings

The use of APP criteria has 'kick-started' a more robust and embedded moderation process, strong support for cross-school and cross-key stage moderation.

APP criteria use has enabled clearer target setting for both teacher and learner

TA has also increased with the use of APP criteria; therefore more accurate summative assessments are being made.

Overall

- APP assessment criteria are starting to support more accurate TAs
- Use of APP criteria have enabled a deeper and broader understanding of a learner's individual needs
- Moderation activities are also being strengthened as a result
- Future considerations should look to increased/further support from LA pilot leaders; encouraging more cross-school and cross-phase moderation; and improving communication of the APP to parent/carers