DISCUSSION PAPER

Moderation of Teacher Judgments in Student Assessment

QUEENSLAND
SCHOOL
CURRICULUM
COUNCIL

Moderation of Teacher Judgments

in Student Assessment

Discussion Paper on Assessment and Reporting

Prepared for the Queensland School Curriculum Council by

Graham S. Maxwell
School of Education
The University of Queensland

Discussion Papers on Assessment and Reporting

1. Common and Different Features of Council and Board Approaches to Assessment and Reporting
2. Teacher Observation in Student Reporting
3. Are Core Learning Outcomes Standards?

These discussion papers are now available at <http://www.qsa.qld.edu.au/research/qscc/index.html#reports>

ISBN 0 7345 2343 2

Queensland School Curriculum Council

PO Box 317, Brisbane Albert Street Q 4002

Level 27, 239 George Street, Brisbane, Queensland, Australia

Telephone: (07) 3237 0794
Facsimile: (07) 3237 1285
Email: inquiries@qscc.qld.edu.au
Website: www.qscc.qld.edu.au

In July 2002, the Queensland School Curriculum Council will amalgamate with the Queensland Board of Senior Secondary School Studies and the Tertiary Entrance Procedures Authority to form the Queensland Studies Authority.

FOREWORD

This discussion paper continues the series of discussion papers on assessment and reporting issues commissioned by the Queensland School Curriculum Council. The intention of the series has been to provoke discussion and to inform policy and practice.

This paper is concerned with moderation — what it means and how it might be practised. The establishment of the new Queensland Studies Authority to replace the Queensland School Curriculum Council, the Board of Senior Secondary School Studies and the Tertiary Entrance Procedures Authority heralds a new era in which old practices will be reviewed and greater coherence will be sought across all stages of formal schooling. This paper looks to the future in suggesting language and concepts for thinking about moderation in a coherent way across the whole school system. In doing so, it recognises that there are different imperatives for moderation at different year levels in the school system and that moderation may serve different purposes and take different forms at different year levels.

A key distinction is made between moderation for accountability and moderation for improvement. The paper suggests that moderation for improvement is needed at all year levels but that moderation for accountability is necessary only where certification is involved and to a lesser extent where teacher judgments are aggregated to provide data for system monitoring. There are suggestions for ways of managing both forms of moderation in future.

This discussion paper is not an official statement of the Council. Rather, it represents the views of the author. It is to be hoped that it will promote useful discussion on an important issue as we move into a new phase of education in Queensland.

J E Tunstall
Director
Queensland School Curriculum Council

June 2002

CONTENTS

INTRODUCTION

Assessment moderation is a process for developing consistency or comparability of assessment judgments across different assessors, programs and schools.

The characteristics of a moderation process will differ depending on the extent to which the assessment judgments are exposed to public scrutiny and comparison.

Two basic forms of moderation can be defined:

- *Moderation for accountability* provides official confirmation of assessments that are reported or used publicly, whether for individual students or for cohorts of students; it therefore involves some external control mechanism or validation requirement.
- *Moderation for improvement* develops the capability of teachers to undertake appropriate assessments and make consistent and comparable judgments but lacks the external control mechanism or validation requirement of moderation for accountability; it therefore involves collaborative processes supporting the professional development of teachers.[1]

Any moderation process can involve a mix of these two orientations. The emphasis placed on each depends on the purpose and use of the assessments and their public visibility and status.

For example, the recording of learning outcomes on an official certificate carries with it the requirement to ensure that the assessment judgments involved are consistent and comparable; otherwise, public confidence in the certificate may be eroded. In this case, the emphasis is on validation — with accountability to the certification authority. However, it is reasonable to expect that the moderation process will also contribute to improvement in assessment practice as well.

At the other extreme, judgments made daily by an individual teacher for formative purposes in the classroom are not, at this stage in the assessment process, being publicly and officially certified and therefore involve much less pressure for verification of their consistency and comparability. In this case, assessment moderation might focus more on assisting the teacher to improve the teacher's assessment practice than to validate the teacher's assessment judgments.

[1] It would be possible — some may say preferable — to distinguish these two processes by different terms (for example, restricting 'moderation' to refer only to accountability-oriented processes and using some other term for improvement-oriented processes). However, by using the single term 'moderation' to refer to both types of processes we can advance towards greater coherence in assessment practice since both types of processes may be appropriate in some form at all levels of the education system. This is discussed later in this paper. Furthermore, the terms 'moderation for accountability' and 'moderation for improvement' mirror existing differentiation between 'testing for accountability' and 'testing for improvement', though in that case the distinction is two purposes or benefits of external tests whereas here the distinction is two different processes or orientations that may or may not co-exist in the same process.

MODERATION FOR ACCOUNTABILITY

Moderation for accountability can be placed within a broader perspective of quality management. There are three basic approaches to the management of quality and they will be referred to here as quality assurance, quality control and quality monitoring as summarised below:

Quality assurance (prior)

- a process for checking that resource and/or delivery standards are in place
- involves accreditation/registration of a person, organisation or program
- focuses on adoption of appropriate resources, expertise and procedures
- by itself, takes a naive view of the effects of context and interpretation
- one example is the AQTF requirements for assessing VET competencies
- another example is approval of teaching plans (work programs) for Senior subjects.

Quality control (during/exit)

- looks at whether outcomes are judged consistently/comparably
- focuses on student performance rather than assessment procedures
- involves verification and confirmation of judgments made
- an example is verification of school-based assessments at end of Year 12.

Quality monitoring (on-going)

- provides feedback for further development of comparability
- may focus on both procedures and outcomes
- may involve authority to require change or improvement (with sanctions)
- examples are monitoring at the end of Year 11 and random sampling after Year 12.

Quality assurance is not in itself a moderation process though it might form part of such a process (as it does for some results on the Senior Certificate). Quality monitoring is essentially moderation for improvement, though it also could form part of a process of moderation for accountability (as it does for some results on the Senior Certificate). That leaves quality control, as defined here, as the definitive representation of moderation for accountability. The following discussion examines how these aspects of quality management are played out in the education system.

The Senior Certificate

Currently, the Senior Certificate is the only official statewide school certificate in Queensland, issued on completion of Year 12. This certificate carries the imprimatur of the issuing authority (until June 2002, the Queensland Board of Senior Secondary School Studies; from July 2002, the Queensland Studies Authority) and serves both as an official state-level attestation to the achievement of the student up to that time and as a basis of access and selection to further stages of education and to work. It is therefore a very serious document for all concerned — students, schools, universities, vocational education and training institutions, employers and the general public. That is, the Senior Certificate is a major point of reference for the interface between schools and the wider community. Its credibility and integrity are crucial to this interface and make moderation for this certificate a major priority for the issuing authority (from July 2002, the Queensland Studies Authority).

Even so, moderation processes for the Senior Certificate are largely restricted to 'Category A' (previously 'Board') subjects. 'Category B' (previously 'Board-registered') subjects do not have an associated process of moderation.[2] Where 'Category B' subjects have embedded vocational education and training competencies, there are quality assurance requirements to be met under the Australian Quality Training Framework (AQTF) but, as suggested earlier, these do not constitute moderation procedures, at least not by themselves.[3] In other words, for the Senior Certificate, moderation procedures are implemented to ensure consistency, comparability and equity for those parts of the certificate where the closest scrutiny will be occur and most serious decisions will be made concerning student achievement, that is, where the consistency and comparability matter most.

For 'Category A' subjects, the set of procedures summarised in the box (next page) is referred to, in totality, as the Senior moderation system. It can be seen that this includes quality assurance and quality monitoring as well as quality control. In the sense in which it used here, quality control is the definitive part of the process, without which the whole set of procedures would not constitute moderation for accountability. For further details see the Senior *Moderation Handbook*.

[2] 'Category A' subjects involve syllabuses that emphasise subject discipline knowledge, academic skills and intellectual enquiry. 'Category B' subjects involve less academic challenge and typically incorporate vocational education competencies.

[3] For a discussion of issues concerning moderation in vocational education and training, see Maxwell (2001b), *Moderation of assessments in vocational education and training*, Queensland Department of Employment and Training <http://www.training.qld.gov.au/reports/vetmoderation/report.pdf>. Greater concern is now being evidenced for the consistency and comparability of judgments of competence for VET qualifications under the AQTF, but practice lags.

THE SENIOR MODERATION PROCESS

For Senior 'Category A' subjects, the following components currently constitute the overall process of moderation (before, during and after the course):

- before: approval of school plans for implementing the subject syllabus
- during: review of each school's assessments through monitoring, verification and confirmation
- after: random sampling (post-hoc evaluation of effectiveness of the moderation process).

At the heart of these moderation procedures is the review process. Monitoring advises schools on the appropriateness of their assessment processes and standards at the end of Year 11 (half way through the course). Verification advises schools towards the end of Year 12 on the appropriateness of their judgments of standards of performance of their students against the exit standards. Confirmation is the final process of validation of the results to appear on the Senior Certificate.

The review panels are the central point of reference. A sample of student folios is sent from each school in each subject to a subject review panel. Each folio includes the relevant pieces of student work or other records of their performances as well as the school's judgments of the standard reached by the student. Teachers' judgments concerning how close each student is to the grade boundaries are also provided. Sample sizes are five for monitoring and nine for verification (or all students if there are fewer than this). Each sample covers the demonstrated range of achievement levels.

Members of each review panel meet and review the sampled folios, considering whether they agree with the teacher judgments of standards demonstrated by each student. In this process they look in the folio for the evidence on which the teacher judgment has been based. Panel members seek agreement within the panel before offering advice to the school. Where adjustments are recommended, discussion and negotiation between the review panel chair the school occurs until a satisfactory resolution is reached.

The following features of Senior panel-based moderation should be noted:

- there is a process for selection and credentialing of panellists (to assure their quality and capability)
- the initial part of the moderation process, the approval of work programs, involves a form of quality assurance
- assessment involves a global judgment, a synthesis of the performance evidence in the student's folio
- schools are given advice mid-course, not just at the end (moderation for improvement)

- only a sample of student folios from each school is considered (unless there are very few students)
- review panels consider whether they agree with the school's judgments of student achievement in full knowledge of those judgments — the focus is on confirming or challenging the school's assessment judgments, not on making an independent or 'blind' judgment to be compared with the school's
- the judgments of student achievement are based on the evidence presented in the folios of student work
- the emphasis is on looking for the evidence to support the school's judgment
- the aim is not to 'test' the school's judgment and report the degree of agreement or disagreement but to reconcile any differences and agree on the results to be awarded

It should also be noted that this assessment and certification process involves a partnership between the central certification authority and each school. School-based assessment imposes a responsibility on each school to undertake assessment professionally and according to principles of assessment endorsed by the central authority. For Senior subjects in Queensland this involves adoption of an overall criteria-and-standards-referenced approach to assessment as well as the specific requirements of each relevant syllabus. The Senior Certificate reports are teacher judgments of the quality of student performance in each subject studied. These judgments are made on the basis of the assessment program adopted within each school, as moderated (verified and confirmed) by the review panels. Each school must therefore conduct its own process of within-school moderation to ensure that there is consistency, comparability and equity of teacher judgments within the school. The review panels are responsible for between-school moderation.

For both within-school and between-school moderation, the focus is on consistency, comparability and equity of judgments within a particular subject. The reference point for this process is the statement of exit standards for levels of achievement in the relevant syllabus. That is, the moderation process focuses on whether the performance of the individual student is appropriately characterised by the level of achievement awarded.

Currently, there is discussion of possible extensions to the range of assessment judgments that might be included on the Senior Certificate. The extent to which such assessment judgments are included in subsequent selection indices and selection decisions will require commensurate attention to their moderation for consistency, comparability and equity.

Public confidence in the Senior Certificate could be eroded if moderation for accountability is not maintained at least at the level of quality currently delivered, with potentially serious consequences for the status and value of the certificate. This does not mean that the current system of moderation is the only one possible or even the most desirable. However, whatever system of moderation is adopted for the Senior Certificate must retain public confidence in the certificate. Queensland is unique in this respect because of its school-based assessment regime and absence of public examinations for its end-of-schooling certificate. This carries with it the responsibility of ensuring that the school-based judgments recorded on the certificate are appropriate and defensible. This requires a process of approval and confirmation of the results — a quality control process.

Pre-Senior certification

A similar level of public importance and visibility for the Senior Certificate does not currently exist for any other school report on student learning outcomes and performance. The Junior Certificate, previously issued on completion of Year 10, was abolished in 1993 (after 15 years of debate) because it was seen as serving no useful public purpose.[4] At the time of its abolition, it had not been supported for some years by any process of moderation. This in itself was indicative of its lack of public importance. Although Education Queensland implemented a Year 10 Statement as a replacement for the Junior Certificate, this report has an even lower public profile and status than the certificate it replaced. That is hardly surprising, but it means that there is currently even less imperative than previously at Year 10 for moderation of the kind supporting the Senior Certificate, that is, moderation directed at verification of the assessment judgments recorded on the Year 10 Statement.[5]

Nevertheless, some formalisation of reporting of the learning progress of students at an earlier point than the end of Year 12 would seem desirable, particularly for those students who leave school or suspend their schooling earlier than the end of Year 12. Such a certificate might include core learning outcomes. If a report were to be issued, some agreement would need to be reached on what form it should take and when it should be issued. As well, some form of moderation for accountability would be necessary to ensure its credibility. The scope of this moderation process would depend on the extent to which such a certificate was likely to inform critical pathway choices and selection decisions.

Even so, the only system-wide accountability-oriented certificate-related moderation that can be considered essential currently concerns the Senior Certificate (and then only for ‘Category A’ subjects). For the moment, at least, this is likely to continue to be the case. At other levels, moderation for accountability currently does not exist so it can hardly be considered essential — merely desirable. As well, even if moderation for accountability were introduced at other levels, it would not involve the same degree of accountability pressure as for the Senior Certificate. What distinguishes the Senior Certificate in this regard is the need for officially ‘signing off’ the results on the certificates.[6]

A further complication here is the different historical legacies for Years 1 to 10 and Years 11 and 12. Whereas for the Senior Certificate there is essentially a single coherent educational framework, common curriculum structure and dominant assessment philosophy, all controlled by a central certification authority, this is not the case for Years 1 to 10. In those years, the separate schools and school authorities have independent responsibilities for

[4] The justification was that there are no serious access decisions at stake at this level and that all students need to obtain additional qualifications beyond Year 10 for worthwhile jobs in the workforce.

[5] In fact, consequences do flow from Year 10 learning outcomes for some students. For example, entry into the Armed Forces requires a ‘pass’ in Year 10 Mathematics — somewhat problematic in terms of what constitutes such a ‘pass’ and how any such judgment is to be arrived at, since the Key Learning Area syllabuses are not tied explicitly to year levels and do not incorporate the concept of ‘pass’.

[6] Vocational education competencies are recorded on the Senior Certificate but are not ‘verified’ by the issuing authority. Each school acts as a Registered Training Authority and must issue its own certificate of attainment as well.

educational frameworks, curriculum structures and assessment approaches, drawing on the centrally produced syllabuses as they see fit. Whether the schools and school authorities would be willing to 'hand back' some of their responsibilities to a central authority for Years 1 to 10 is unclear. Probably, that would happen only if the authority developed worthwhile certificates incorporating an agreed assessment framework and reporting structure.

School reports

Individual schools issue their own reports, typically each semester, on their students' progress. There currently is no standard form that these reports might take. That is, there is no basis, and currently it would seem no demand, for common reporting across schools, and therefore no basis on which such reports might be moderated. Nevertheless, it is important for the school to issue reports that are consistent within the school context. This means that schools ought to have an internal moderation process. This in turn requires coherent school assessment policies and procedures. Moderation within a school needs to link both accountability and improvement. The accountability may be limited to the school context. In some cases, there would be benefit in extending these moderation processes to include other schools, for example in a district or a diocese (or even, voluntarily, across sectors).

It should be noted that external moderation depends on within-school moderation occurring first. Unless there is coherence and consistency within a school, it is impossible to achieve coherence and consistency between that school and other schools.

MODERATION FOR ACCOUNTABILITY: OTHER ISSUES

Three other issues relating to the pre-Senior years should be mentioned: system monitoring; Education Queensland's Rich Tasks and reporting junctures; and Education Queensland's Year 2 Diagnostic Net.

System monitoring

There is no common and agreed system for reporting learning outcomes for Years 1 to 10 across the state. The Key Learning Area (KLA) syllabuses provide a potential framework for monitoring student progress through sequences of core learning outcomes (CLOs). However, there are currently no requirements on schools concerning the ways in which they might implement these syllabuses and the forms of recording and reporting they might employ. Assessment moderation, whether accountability-oriented or improvement-oriented, requires some common framework of standards against which teacher judgments of student performance can be compared.[7]

Education Queensland's *Task Force on Assessment and Reporting* (EQ 2002) recommended that government schools negotiate with their school community on their forms of reporting on student progress but that schools should at least record their students' progress on the core learning outcomes of the KLA syllabuses for Years 1 to 10. This recommendation makes a critical distinction between recording and reporting. If this recommendation is implemented, it will mean that a common framework exists to enable implementation of moderation procedures. Collation of these school databases into district, like-school and system databases and profiles would require attention to the consistency and comparability of the data — there is not much point in collating data lacking consistency and comparability. That is, some form of moderation to ensure the consistency and comparability of the data across the system would be desirable and necessary. The same could be said for the collation of data across schools in any educational jurisdiction (for example, across a Catholic Diocese). Even so, the degree of moderation thought necessary would depend on how public and important the comparisons might become: the more public and the more important they become, the less tolerance for ambiguity there would be and the more need for some form of moderation for accountability.[8]

[7] The sequences of core learning outcomes for the KLA syllabus may provide such a framework, at least for recording student progress, and therefore as a basis for moderation of standards (see the previous paper in this series — Maxwell, 2002). However, core learning outcomes will need further development to function as a satisfactory framework of standards for recording student progress. Furthermore, they do not represent all the dimensions of student learning that should be assessed and reported, merely the 'core'.

[8] This would probably be self correcting. That is, the extent of public dissatisfaction with the data would determine whether the government allocated additional resources for improved moderation. However, there is always the danger that the public would demand (simplistically) other forms of data, such as test data, instead.

A case can be made for moving in the direction of system monitoring that is founded on school-based teacher judgments rather than external standardised tests. School-based teacher judgments offer the possibility of richer, more authentic, more sensitive and more comprehensive assessments of student progress as compared to standardised tests. They also allow assessments to be devised according to modern understandings of knowledge and learning, especially in terms of developing capacities for life long learning,[9] developing complex knowledge structures and repertoires, developing personally meaningful and useable knowledge, developing generalisable and transferable knowledge, and recognising the effects of context on performance. All of these require assessment that is more naturalistic than standardised tests can be, especially involving more authentic, holistic, contextual and meaningful activities, that is, involving school-based teacher judgments. Clearly, if such judgments are to be valued for system monitoring, as they should be, then a system of moderation for accountability will be necessary.

Here, a caveat is in order. Even though this form of moderation can be termed 'moderation for accountability', there are important differences between moderation for reporting and certifying individual performance — where a high level of comparability is demanded — and moderation for reporting and analysing group performance — where a lower level of individual comparability can be acceptable. The level of acceptable comparability depends on public perceptions of the veracity of the assessment data and the use to which the assessment data are put — the more public or serious the consequences, whether for individuals, schools or systems, the higher the level of comparability demanded.

Year 2 Diagnostic Net

The Year 2 Diagnostic Net is a special case of a fairly high-stakes assessment conducted in the early years of schooling. The intention is to identify students who are at risk of falling behind in their literacy and numeracy development and to provide additional resources to schools to assist these students to make better progress.[10] This does not result in a public report on each student. However, it is clearly important to make appropriate and consistent identification of the students at risk, both from the point of view of the students and their parents and from the point of view of the schools and school authority.

All stakeholders, particularly the students and their parents, have an interest in ensuring that those who should be identified as being at risk are identified as such. Also, all stakeholders, particularly government as the funding agency, have an interest in ensuring that those who should not be identified as being at risk are not identified as such. The tendency would be to err on the side of the former rather than the latter since receiving additional support when you do not need it is less serious than not receiving support if you need it. However, over-identification of those at risk could spread the available resources more thinly with serious consequences for those most in need. In other words, these assessments involve fairly high-

[9] KLA syllabuses posit the following attributes of the life-long learner: a knowledgeable person with deep understanding; a complex thinker; a creative person; an active investigator; an effective communicator; a participant in an interdependent world; a reflective and self-directed learner. Other ways of representing such 'personal learnings' can been devised and could be adopted (Cumming, 2002). Such learnings are difficult, even impossible, to assess on standardised tests but are observable in classroom contexts.

[10] The three dimensions are reading, writing and number.

stakes selection decisions and it is important to make consistent, comparable and equitable judgments.

The process involves development of an Individual Student Profile together with the application of specially designed validation tasks to selected students. The validation tasks are seen as assisting teachers in confirming their other observations and judgments. Schools are also encouraged to develop moderation processes involving teachers within the school and across school clusters to confirm the teacher judgments.

The Year 2 Diagnostic Net is generally perceived as a successful process, though there are lingering workload issues and important substantive issues (see, for example, van Kraayenoord and Luke, 2000). Some of the components that contribute positively include:

- identification of specific 'tasks' or 'questions' for students
- explicit developmental continua against which students are assessed
- explicit criteria for making judgments of the levels on these continua
- one teacher responsible for coordination and review of student folios with each school
- a moderation meeting of teachers across schools in the district one day each year
- aiming for common interpretation of the continua across the state
- seeing the aim as being professional development as well as consistency of judgment.

In the moderation meetings, teachers compare their identification of students at risk and contest and verify each others' judgments. Where there is an unresolvable difference, the case goes to an external arbitrator. The moderation process is seen as giving advice to schools concerning their judgments. Schools take this advice and make their final judgments. It is reported that teachers have gradually acquired more confidence to challenge each others' judgments 'more genuinely'.

The Year 2 Diagnostic Net is a good example of a duality of moderation for accountability and moderation for improvement being intertwined. Each motivation benefits the other. The teacher judgments are 'tested' and validated but the process also contributes to, and in turn depends on, professional development of the teachers involved. This is a never-ending process since there are always new teachers becoming involved and new student cases to consider.

New Basics and Rich Tasks

Education Queensland's New Basics trial includes Rich Tasks as the assessable outcomes. The assessment framework for Rich Tasks involves merit ratings (grades) and standards rubrics defined in terms of 'indicative standards descriptors'. As for Senior subjects, the approach taken is to assess the standard of the student work by matching it to the standards descriptors and making a judgment of its quality (see New Basics Branch 2001). A process of moderation will be directed at providing opportunities for sharing among teachers, consulting and negotiating on the application of the standards descriptors and ratifying teachers' judgments. Because of the high profile of this project there are elements here of moderation for accountability as well as accountability for improvement.

MODERATION FOR IMPROVEMENT: THE CURRENT SITUATION

Even without a strong accountability requirement for verifying teacher's assessment judgments, there are other reasons why some form of moderation might be desirable at all levels of the education system. This can be referred to as moderation for improvement in the sense that it is directed at enhancing the quality of schools' assessment programs and teachers' judgments of student progress. In other words, this form of moderation is explicitly linked to professional development and directed at overall improvement in the quality of the education system.

Currently, moderation for improvement is accidental rather than deliberate at all levels of the education system. Where there is moderation for accountability, backwash effects on schools ensure that there are some consequences that lead to gradual improvement in the quality of assessment practice and teacher judgment. Even so, in this case moderation for accountability is the primary consideration and any improvement in the quality of assessment practice and teacher judgment is an incidental consequence rather than a deliberate expectation. This is especially so for the Senior Certificate where the formal moderation system operates as a process conducted between a school and the central authority (represented by a review panel). Improvement results from feedback to the school from the panel. There is no formal process of sharing across schools.[11]

Education Queensland's Rich Tasks and Year 2 Diagnostic Net both allow a more deliberate sharing across schools and more deliberate encouragement of improvement through this process. In fact, accountability and improvement are in these cases explicitly intertwined. However, these are special cases that do not affect all teachers and all levels.

It is not difficult to initiate a process of moderation for improvement. All it needs is some form of conversation between teachers focused on their assessment practice and assessment judgments. This could be informal and voluntary. Alternatively, it could be more formal and obligatory. The former is likely to be more successful in the absence of any public use of the data demanding moderation for accountability as well.

The following reasons can be advanced for the importance of moderation for improvement:

- there is a need for greater coherence in reporting of learning outcomes across the system
- recent research has highlighted serious deficiencies in teachers' assessment practices[12]
- teachers' judgment ought to be the basis for monitoring of the system performance
- making more use of teachers' judgments requires that they be made more dependable
- dependable judgments require common understandings and agreements on standards

[11] This would seem to be one reason why some teachers ask to attend review panel meetings as observers. However, this is not a substitute for a process of improvement-oriented moderation across schools.

[12] Lingard, Mills, Bahr, Chant and Warry 2001, *The Queensland School Reform Longitudinal Study (QSRLS)*, State of Queensland, Brisbane.

- dependable judgments require confirmation that teacher judgments are comparable
- confirmation of comparability requires sharing and contesting of teacher judgments
- focusing on student work is known to produce the most powerful professional development
- improved coherence and quality across the system needs to grow out of professional activity
- improvement is a progressive and never ending process of 'growing the culture'.[13]

All of these reasons suggest that it would be desirable to implement some form of moderation for improvement at all levels of the system. What this might look like is considered later. However, it should be noted that there was an attempt in the early 1990s to introduce moderation for improvement in Queensland. This was in association with the Student Performance Standards which were associated with the National Reporting Framework. This attempt, and in fact the whole framework, is thought to have failed for the following reasons:

- the workload involved, the speed of implementation and the lack of preparation
- poor alignment between teaching/learning activities and national reporting framework
- lack of understanding of the national reporting framework and its implications
- teachers were not used to or prepared for the portfolio approach that was needed.

This is a cautionary tale though the circumstances differ now. It also must be noted that many Catholic Dioceses have been making steady and substantial progress in implementing KLA syllabuses, supported by processes of across-school moderation. It is clearly possible to do this.

[13] Any process of educational change that does not enlist the professional involvement and commitment of teachers has been found in the past to have a very high probability of failing.

WHAT IS MODERATION ABOUT?

Merit versus milestone standards

Moderation, in the way it is conceived in this paper, is concerned with the consistency, comparability and equity of professional judgments about the performance levels demonstrated by students. In making any such judgment, the assessor needs to collect evidence on performance and to interpret it by reference to some framework of standards. This reference framework may involve 'merit' standards or 'milestone' standards. 'Merit' standards represent a set of ordered categories showing different levels of quality in student performance, typically in relation to the expectations of a particular course of study, and can be thought of as 'quality ratings'. 'Milestone' standards also represent a set of ordered categories but in this case showing different levels of progress from 'initial' to 'more advanced' status as a learner, and can be thought of a 'progress markers'.[14]

Merit standards or quality ratings are a more traditional way of representing standards and reporting student learning. For example, in Senior subjects, 'exit standards' are merit standards or quality ratings, referred to as Levels of Achievement and indicated by five ordered categories: Very Limited Achievement (VLA), Limited Achievement (LA), Sound Achievement (SA), High Achievement (HA) and Very High Achievement (VHA). On the other hand, KLA syllabuses have adopted milestone standards or progress markers, which could be termed Levels of Development, involving six steps along sequences of learning outcomes. These learning outcomes are core learning outcomes and do not represent all the learning outcomes that students ought to demonstrate. However, the KLA syllabuses are silent on the question of how these additional learning outcomes might be represented as assessable standards and whether they should be represented by quality ratings or by progress markers.[15]

Quality ratings and progress markers each have advantages and disadvantages. Quality ratings have the advantage of familiarity and they fit the presumption (and experience) that at any point in time (such as the end of a course) there will be a spread of quality in student performance. Such ratings arise from the assumption or expectation that schooling involves competition and that assessment should explicitly report the demonstrated range of quality in students' achievement. Even if we specify and 'fix' the standards in advance as targets for student learning, there is an expectation that some students will do better than others and that some students may fall well short of the ideal (perhaps 'fail' or fall short of 'satisfactory').

[14] This is not to say that 'progress markers' do not incorporate 'increasing quality'. Progress encompasses many characteristics of progression from novice to expert. This can be characterised more descriptively in terms of increasing elaboration or complexity or it can be characterised more interpretively in terms of relative quality or merit. A novice-to-expert continuum emphasises internal comparison within the student, allowing continuing celebration of their own 'personal best' and idiosyncratic representation of their own 'personal profile'. However, at any time, student positions on a developmental continuum can be compared with those of other students (creating comparative distributions and rank orders) or with benchmark expectations (creating grades or ratings). It is not a matter of either-or but which is to be foregrounded — that is, which is to provide the primary basis for our discourse with students about their own learning.

[15] For further discussion of the 'merit' versus 'milestone' similarities and differences in relation to Senior and KLA syllabuses, see the first paper in this series (Maxwell 2001a).

On the other hand, progress markers, such as KLA core learning outcomes, have the advantage of representing a developmental continuum which allows progress to be explicitly represented from year to year. Progress markers are expressed positively to indicate the progress that the student is making. With quality ratings, it is difficult to avoid the implication that grades lower than the top grade fall short of the ideal (are deficient in some respects).[16]

There is often a need to consider how well the individual student is progressing in relation to other students. This is somewhat transparent in the case of quality ratings since comparison is built into the ratings. With milestone markers, comparative data could be generated from the individual student data by profiling the class, the school, the district/diocese or the state and comparisons made in varieties of ways. Alternatively or additionally, 'reasonable targets' could be set for particular ages or years and the individual student compared with those.

Whether merit or milestone standards are adopted, the principles of moderation are essentially the same. In both cases, the judgment being made is whether student performance, as evidenced by performance on a single task or a collection of tasks (for example, in a portfolio of student work), is an instance of a particular standard along a sequence of standards (or, alternatively but equivalently, which of the standards along the sequence it fits). Moderation is about confirming this judgment.

Moderation versus scaling

Linn (1996) referred to moderation involving human judgments as 'social moderation'. He contrasted this with 'statistical moderation' where adjustments are made to sets of scores to make them comparable or equivalent in some way. An example of statistical moderation is scaling a school's internal assessment scores in a subject against the distribution of external examination scores for that group of students in that subject, as occurs in New South Wales with Higher School Certificate subject results.

Statistical adjustment of this kind is best referred to as scaling, not as moderation. In fact, this form of statistical scaling is not moderation in the sense adopted in this paper since it does not ensure that the individual student results represent comparable and interpretable standards of performance in the subject. Since the two sets of marks refer to different assessment performances, the performances may not reference the same underlying characteristics. Furthermore, even if they did reference the same characteristics (for example, where the school assessments mirror the external assessments in their content and form), the standard of the performances could be quite different. Therefore, the scaling merely realises an

[16] Quality ratings may be represented by specified standards, as they are in Senior syllabuses. However, these standards are developed on the basis of expectations and experience about the range of quality in student performance likely to be produced. There is a kind of Catch 22 about this for some students: They are offered a target for their learning but it is actually beyond them (at least within the allowed timeframe). This is also somewhat like the joke about not being able to reach an intended destination if you start from where you are at the moment. With the best will in the world and the best teaching in the world, some students cannot reach the top grade — and the system is set up so that this is the case. Nevertheless, where quality ratings are represented as standards (through description and exemplification), they do offer clear referents for assessing the quality of student performance, clear targets towards which student learning can be advanced and clear expressions of the possible range of differentiated performance. This also encourages the discourse on assessment to be primarily about intrinsic characteristics of performance and how to reach towards higher quality rather than about extrinsic labels such as points, marks and grades.

expectation that one set of results mirror the other set of results. The process of scaling adopts the assumption of equivalence but cannot verify it.

The term ‘social moderation’ has acquired some currency in Queensland as referring to what is here called ‘moderation for improvement’. This allows a distinction to be drawn between ‘panel moderation’ (bureaucratic) and ‘peer moderation’ (social) in the current context and is probably therefore a reasonable terminology to adopt. However, it contradicts the accepted international usage of ‘social moderation’ as ‘moderation involving human judgments’. In this sense, current moderation for the Senior Certificate is also ‘social moderation’. It seems better to make the distinction between moderation for accountability and moderation for improvement — that is, a distinction of purpose rather than style.

It must be emphasised that the scaling procedures in Queensland for calculating tertiary entrance ranks (Overall Positions) are not about moderation. These scaling procedures are concerned with comparing performance in one subject against performance in other subjects in order to calculate a general index of overall performance. That is, this kind of scaling is for calculating a fair measure of overall performance whatever combination of subjects students have studied, not for ‘moderating’ subject performance across schools. Another way of saying this is that this kind of scaling is a between-subjects adjustment to take account of different student clientele in different subjects *when an overall index of relative performance must be calculated* whereas moderation is directed at confirmation of within-subject judgments of student performance against defined standards.

Consistency, comparability and equity

Three principles relevant to moderation are consistency, comparability and equity.

Consistency

Consistency appears to be a broader term than comparability, although comparability is the issue of major concern for the Senior Certificate. Consistency involves constancy of judgment by the individual teacher with respect to the same evidence at different times — if the same evidence was looked at again, the same judgment would be made. Consistency also involves equivalent application of standards across different types of evidence and different opportunities for assessment — this applies to the same student and to different students.

In addition to within-teacher consistency, we need between-teacher consistency — both between teachers within the same school and between teachers in different schools. This implies that teachers are interpreting and applying standards in equivalent ways so that they would confirm each other’s judgments about the standards demonstrated by their students.

When certification is involved, all of these types of consistency are important, though the public focus is usually on between-school consistency. Within-school consistency, including teacher consistency, is often assumed to be unproblematic or the responsibility of the school rather than the certification authority. Even so, it is impossible to have between-school comparability without within-school (and within-teacher) consistency so the latter is a necessary foundation of moderation for accountability.[17]

Comparability

Comparability has been the term adopted in Senior Certificate moderation for a similar concept. Here, the starting point has been the need to ensure that each student's Level of Achievement in a subject, as recorded on the Senior Certificate, is referenced to a common performance standard, that is, indicates equivalent performance to that of other students awarded the same result. This is a within-subject comparison against the performance standards for the subject.[18]

An important feature of this representation of comparability is that the focus is on the assessable performance, not on the assessment task. It is not necessary to have a common assessment task or test in order to establish comparability. Students can be set different tasks or tests but demonstrate a common standard of performance. Performances can differ in their surface features but reveal equivalent levels of knowledge, understanding and skill. Further, this does not require identical aspects of knowledge, understanding and skill but equivalence of standard in terms of the *characteristics* of the knowledge, understanding and skill expected for that level of achievement.

In fact, in Senior Certificate moderation, assessment tasks (and their timing and context) differ across schools. Each school implements a subject syllabus differently within the general framework constraints of the syllabus and according to their approved teaching plan ('work program'). The contents of student folios therefore differ. Yet, review panels have no conceptual difficulty in focusing on the underlying characteristics of the student's work and its relationship to the defined performance standards.

This can be enacted at all levels in the assessment process. Just as students in different schools do not need to undertake the same tasks for comparability of performance standard to be established, so too students within the same school do not need to undertake common tasks or tests for comparability of performance standard to be established. Although not yet common, students could pursue individually different learning programs and assessment programs and still be judged against common performance standards. As in the between-school situation, comparability is achieved by comparison with the performance standards, not by using common assessment tasks.

Equity

Equity in this context can be defined as the opportunity for every student to demonstrate their current capability. Opportunity can be idiosyncratic. In a standards-referenced system, assessment judgments primarily involve comparison of the student's performance relative to the standards and only secondarily relative to other students. Since any standard can be

[17] Further elaboration of teacher consistency in assessment judgments can be found in Curriculum Corporation 2000, Queensland School Curriculum Council 2000, and Queensland School Curriculum Council 2002 (see references at the end of this paper).

[18] Although there is a general notion of cognitive demand and intellectual challenge for all 'Category A' subjects, standards for levels of achievement necessarily depend on the internal logic of the subject itself, which includes the way in which the characteristics of the subject are conceptualised.

demonstrated in a variety of ways, the critical concern ought to be whether students have had adequate opportunity to demonstrate what they know and can do. Standard assessment requirements, such as common assessment tasks, do not necessarily enable each student to perform optimally. Adaptation to the current state of development of the individual student is required for a truly equitable assessment process. Characteristics of the task and the context are therefore of critical importance in interpreting the student's performance and judging the standard reached. Moderation buttresses equity by checking that these characteristics have been properly considered in interpreting the evidence and that the student's performance has been appropriately compared with the standard.

Comparability of assessment judgments means that there is agreement that the assessed performances are appropriately classified in terms of the standard they demonstrate. This involves both similar interpretation of the standards and similar recognition of performances that demonstrate those standards. A moderation process is therefore one involving approval of assessor judgments, with the implication that there may need to be some adjustment of those judgments to conform to the common standard. It is not a passive process that simply checks how much agreement there is. It is an active process in which assessment judgments are aligned with each other to create consistency of interpretation and implementation of standards across the whole system. The intention of any moderation process is to resolve any differences of opinion rather than to calculate and accept the degree of disagreement.

MODERATION PROCESSES

Moderation processes, whether for accountability or improvement, are directed at supporting and confirming assessment understandings and assessment judgments.

Assessment understandings

Assessment understandings can differ even when there is a clear statement of standards (whether milestone standards or merit standards) and some relevant exemplars of those standards (examples of student performance representing each standard). This is necessarily so because both require interpretation and integration into the mental schema of the individual assessor. Meaning is not transparent; rather it must be constructed. Standards are conceptual categories. The words used in the statement of a standard denote the standard, that is, attempt to convey the meaning of the standard. Whether the same meaning is derived by each assessor is problematic. Similarly, exemplars typify the standard but a process of induction is necessary to arrive at the standard itself. Other instances of the standard may differ in unpredictable ways from these exemplars but still be equivalent realisations of the standard.

Further, standards are fuzzy categories, in the sense that they are broad classifications with imprecise boundaries. This is an advantage in that it allows some tolerance for difference of interpretation while encouraging sufficient convergence for practical implementation. The degree of tolerance allowed and the degree of convergence desired depends on the circumstances and the use to which the assessments are put but there are limits to what is reasonable to expect. There is an inherent amount of ambiguity involved. In attempting to reduce the amount of ambiguity the cost/benefit ratio increases exponentially.

Some degree of convergence of interpretation of standards requires transactions of understandings among assessors. That is, only by talking to each other can assessors come to appreciate whether and how their understandings may differ from those of other assessors and only through further discussion can assessors bring their understandings into greater alignment. Thus, the foundation for any moderation system needs to be discussions among assessors (teachers).

Discussions among teachers about standards can occur before, during and after assessment. They also need to be ongoing. There is never a point at which such discussions can stop since there are always new teachers to the system, new teachers to the school, new teaching assignments, new syllabus revisions, new assessment ideas, new teaching plans, new social contexts and new generations of students. All of these require rediscovery and rethinking of the performance standards.

Such discussions need to start with the syllabus statements of standards, the criteria for representing those standards, and any available exemplars. The question to be asked is what does each assessor interpret each standard to mean. This requires 'think aloud' strategies and conversations among teachers in order to challenge and synthesise their understandings. A good starting point is to ask what makes each exemplar representative of a particular standard (and whether it is). This can lead to discussions about the design of assessment tasks and opportunities

— especially whether particular tasks allow for demonstration of relevant standards and how variation in the context may affect the meaning of the student's performance.[19]

Assessment judgments

Assessment judgments can differ even when there is substantial convergence of understandings among teachers of the meaning of the standards. Judgments about standards are necessarily qualitative, that is, require direct apprehension by thoughtful consideration of the characteristics of the performance. Such judgments necessarily involve:

- fuzzy criteria rather than sharp criteria
- multiple interrelated characteristics rather than single isolated characteristics
- defensible justification rather than correct decision.[20]

For convergence of the assessment judgments, not just the understandings of standards, some opportunity is needed for comparison with the judgments of other teachers. Feedback from review panels can provide this, especially if there is opportunity to answer any queries and negotiate any differences of viewpoint, as happens with Senior moderation at present. However, this could also happen more directly, at least with improvement-oriented moderation, through interchanges between teachers. In fact, having teachers focus on their judgments of the standards demonstrated by student performance — that is, real examples from their classrooms — is the most powerful form of professional development available. It would seem desirable that, even where there is a panel-based accountability-oriented moderation process, direct exchanges and challenges between teachers across the school system also be encouraged because of their potential to raise the quality of assessment practice and consistency of assessment judgments.

Whether teacher judgments are confirmed through panel review or teacher interchanges, it is impossible to undertake a complete census of all students. Sampling is necessary. For panel review, it is desirable to have systematic representative sampling. For teacher interchanges this is less necessary and there might be some advantage in teachers sharing puzzling or difficult cases as well as more straightforward cases.

For this process of judgment confirmation to proceed, the following components are needed:

- the student's work (which would typically be a portfolio of evidence)
- the characteristics of the context (including the circumstances of the student's performances relevant for interpreting the nature of the student's own contribution)

[19] For example, providing assistance where necessary for a student to make progress with a task may be a necessary and desirable strategy — in the sense proposed by Vygotsky, helping the student to bridge their zone of proximal development — but implies a different interpretation of the performance than if it was done independently. Knowing the context is essential to interpreting the performance.

[20] See Sadler (1986). The distinction between 'defensibility' rather than 'correctness' is a key point. There is no objective gold standard for such judgments. Different assessors can agree on a match between performance and standard but for different reasons. This is of some interest for further discussion between the assessors but in a practical sense the achievement of agreement between assessors on which standard applies is sufficient. What really matters, though, is whether each can justify their judgment cogently, coherently and convincingly (to other assessors and to the student).

- the characteristics of the assessment tasks or assessment opportunities (including what the student was asked to do or given an opportunity to do — the demand features)
- the assessor's interpretations of the characteristics of the performance(s) in terms of expressed criteria or indicators
- the assessor's judgment of the standard demonstrated (with their justifications).

The assessor's interpretations and justifications might be delivered orally rather than in writing. The confirmation process then can be conceived as a professional conversation. One assessor 'puts the case' for the judgment and the other side considers whether they agree. That is, the confirmation process is conducted in full knowledge of the assessor's judgment and the justification for it. Independent (or 'blind') review may be useful for research purposes but tends to create divergence rather than convergence. It is better to model the process on 'judicial' review. If the assessor's judgment of the standard cannot be supported, the process needs to move to 'reconciliation' mode where a resolution of the differences in interpretation and judgment are explored through discussion and negotiation.

How far this proceeds depends on whether there is an accountability orientation or an improvement orientation. In some case the parties might simply agree to differ. In other cases, as with Senior review panels, one side may have the ultimate power to force the issue, though presumably not until all opportunities for negotiation have been exhausted.

CONCLUSIONS

Essentially, the two forms of moderation discussed in this paper — moderation for accountability and moderation for improvement — are differentiated by whether they service official reporting or professional development. Official reporting demands some form of moderation if the assessments are to have credibility. The degree of importance attached to decisions based on the official reports determines the strength of moderation thought necessary. The Senior Certificate is a public document and it contributes to serious decisions about access to future studies. Hence, it requires substantial moderation.

Even so, parts of the Senior Certificate are not moderated because they are considered to have less importance in terms of selection decisions by tertiary institutions. For these parts of the certificate, comparability is either presumed to exist or to be of less consequence. It may be of less consequence if quotas are under-subscribed rather than over-subscribed. Where quotas are over-subscribed, selection decisions involve preference for one student over another and this requires that the assessments on which those selection decisions are based have some degree of comparability.

The Senior Certificate stands on its own in this respect and its credibility must continue to be supported by a system of moderation such as currently exists. This system might benefit from review in terms of its details but the need for such a system and its adequate resourcing cannot be denied. In fact, if it were decided that additional components should contribute more directly to high-stakes selection decisions then additional resources would be needed to ensure their integrity through moderation.

No other assessments across the education system currently have the same importance as those for the Senior Certificate. This is unlikely to change in future. If some form of certification were devised for other stages of education — for example, at Years 3, 6 and 9 — this would not match the Senior Certificate in terms of seriousness of consequences. While some form of moderation for accountability would be necessary and this would have resource implications, it would not need the same level of confirmation. Any allocation of resources in this direction would need to be additional to those needed to sustain Senior moderation.

A similar situation would exist if the focus was on system monitoring rather than individual reporting. This might provide a firmer justification for investing resources in moderation at these levels. A case can be made for building school and system profiles by aggregation from the database of teacher judgments than by standardised testing. It can be argued that this would provide more valid and more comprehensive representation of system performance. Necessarily, public credibility in the data would require some form of moderation.[21]

Whether or not there is moderation for accountability, moderation for improvement is essential for developing coherence and consistency across the educational system. It also offers the most powerful form of professional development. It has been argued in this paper that coherence and consistency in the development and application of standards is not possible without professional conversations among teachers. In the absence of certification requirements, this would need to be driven by management expectations on schools and

[21] For an extensive exploration of this issue see Pellegrino, Chudowski and Glaser (2001).

teachers, that is, as part of the cultural definition of what schools and teachers do. This kind of cultural change does not happen quickly but needs progressive encouragement over time.

The essence of moderation for improvement is conversations among teachers about their assessment practice, interpretations and judgments. Various suggestions already exist for what this might involve.[22] A start could be made by encouraging teachers to pair themselves with another teacher and beginning a professional conversation about assessment. Some pairs might be within-school and some might be between-school. Some might be within a particular key learning area and year level and some might involve other ley learning areas and other year levels. These pairings could be changed from year to year to provide fresh perspectives. Some could grow into larger groups and some might become more formalised. The process itself can evolve. It is more important simply to begin.

[22] For example, see QSCC 2002, *Position paper and guidelines: An Outcomes Approach to Assessment and Reporting, <http://www.qsa.qld.edu.au/research/qscc/pdf/PositGLdoc.pdf>.*

REFERENCES

Cumming, J.J. 2002, *A Framework for Assessing and Reporting Personal Learning*, Paper Presented at the Second Conference of the Association of Commonwealth Examination and Accreditation Bodies, Malta.

Curriculum Corporation 2000, *Consistency of teacher judgment CD-ROM: A training and developmental resource*, Curriculum Corporation, Carlton, Victoria.

Education Queensland 2002, *Report of the Task Force on Assessment and Reporting*, Education Queensland, Brisbane.

Lingard, R.L., Mills, M.D., Bahr, M.P., Chant, D.C. & Warry, M. 2001, *The Queensland School Reform Longitudinal Study (QSRLS)*, State of Queensland, Brisbane.

Linn, R.L. 1996, Linking results of distinct assessment, in M.B. Kane & R. Mitchell (Eds), *Implementing performance assessment: Promises, problems and challenges* (pp. 91—105), Lawrence Erlbaum, Hillsdale, New Jersey.

Maxwell, G.S. 2001a, *Common and different approaches to assessment and reporting in Council and Board syllabuses*, Queensland School Curriculum Council, Brisbane.

Maxwell, G.S. 2001b, *Moderation of assessments in vocational education and training*, Department of Training, Brisbane, <http://www.training.qld.gov.au/reports/vetmoderation/report.pdf>.

Maxwell, G.S. 2002, *Are core learning outcomes 'standards'?*, Queensland School Curriculum Council, Brisbane.

New Basics Branch 2001, *New Basics: The why, what, how and when of Rich Tasks*, New Basics Branch, Education Queensland, Brisbane.

Pellegrino, J., Chudowski, N. & Glaser, R. (Eds) 2001, *Knowing what students know: The science and design of educational assessment*, National Academy Press, Washington D.C.

Queensland School Curriculum Council (QSCC) 2000, *Consistency of teacher judgment: Research report*, QSCC, Brisbane.

Queensland School Curriculum Council (QSCC) 2002, *Position paper and guidelines on assessment and reporting for years 1 to 10*, QSCC, Brisbane.

Sadler, D.R. 1986, *Subjectivity, objectivity and teachers' qualitative judgments*, Discussion Paper 5, Board of Secondary School Studies, Brisbane, <http://www.qsa.qld.edu.au/yrs11_12/assessment/discussionpapers.html>.