

Realising and releasing potential 40 years on

Gabrielle Matters, Australian Council for Educational Research
Claire Wyatt-Smith, Griffith University, Brisbane

One of the most significant differences in assessment arrangements across the world is in the underpinnings of assessment and standards and, more specifically, understandings about the nature and function of standards in relation to classroom practice, assessment techniques and accountability expectations. And one of the defining characteristics of a jurisdiction's assessment arrangements is the nature of the regime that sets and marks the components of an assessment program for which student results appear on a certificate. There are many possible variations.

The exercise of looking back after forty years is the stimulus for this paper, with gaze focused on the trajectory of a system—the Queensland model of externally moderated school-based assessment in a high-stakes environment (the pre-tertiary year). We intend the paper to provoke re-thinking about key issues related to the topic of school-based assessment more generally, including definitions applied and discussions held about the practices for achieving validity, reliability and equity. Presently in Australia, such re-thinking has immediate significance given the changing educational, political and social contexts in which assessment regimes are being enacted and reconsidered. Further, given that the Queensland system is sufficiently distinctive from other forms of standards-based assessment internationally, it merits sustained and systematic scrutiny.

In what follows, first we present an historical perspective on the Queensland model and examine the roads taken and those not taken. Of interest is the set of circumstances where a system was subjected to forces that tested it and made it change into a model of standards-based¹ assessment that has enjoyed high levels of confidence from the public and the profession. We characterise the various identifiable eras in the development of the model and consider the opportunities and challenges taken up to date and those emerging. In looking backwards, we describe the evolution of the Queensland senior system through four eras, defining its key elements. In looking forwards, we propose ways for strengthening connections between the original conceptualisation of the system and current practices, maintaining as the centrepiece the commitment to rigorous assessment.

Setting the frame: Forty years ago and forty years on

Forty years ago it was 1968. Five years had passed since the assassination of President John Kennedy in the United States. The 16-year-long Vietnam War (Second Indochina War) was barely half-over. May 1968 saw a series of student protests and a general strike that caused the eventual collapse of the De Gaulle government in France. Australian Prime Minister Holt had drowned in the surf at Cheviot Beach just before Christmas the previous year. The new Premier of Queensland was Johannes Bjelke-Petersen, leader of an arch-conservative political party. And Queensland schools and universities were reeling from the aftermath of the 1967 Senior Physics paper — only 30 per cent of candidates had 'passed'² (Clarke, 1987). (Appendix 1 contains more information about this episode.)

¹The terms standards-based and standards-referenced are sometimes used interchangeably. Assessment with pre-determined standards is not the same as assessment with standards based on 'cut scores' established post hoc. See discussion later in this paper. For clarity of message here, we reserve the term standards-based for the situation where verbal descriptors exist before assessments are set or marked or graded.

² Received a letter-grade of C or better (i.e. 'passed', in the now outdated language)

In response to public dissatisfaction, the government set up a review of public examinations. It was chaired by Dr William Radford, then director of the Australian Council for Educational Research (ACER), who stated (Radford, 1970):

Examinations, and particularly what we in Australia call 'public' examinations, have been under fire for decades. Doubt has been cast on them on social grounds as well as technical. No matter who controls them or however humanely used, they have been accused of serving the status quo in any society by sifting out at various levels a limited number for whom alone the benefits of further education or of employment of high status are then available. No matter how well prepared or objectively marked (and many public examinations are excellent in both respects), the results obtained from them are criticised both because they represent performance at a single point in time, and are therefore subject to the known vagaries of mental functioning, and because, being limited in scope by the time available, they rarely if ever sample adequately all the knowledge, skills, abilities, understandings, interests, attitudes, and habits that form the complex objectives of a curriculum or even of a course of study in a single subject.

The government accepted the recommendations of the Radford Review. Public examinations were abolished. The last Senior Public Examination in Queensland was held in 1972. To this day, Queensland and the ACT are the only states in Australia where no external examinations occur in the senior years of schooling and where there has been a focus on teacher professional judgment and the complementarity of formative and summative assessment. A common question, even after 35 years, is: How is it that a right-wing government took the radical step of abolishing examinations? We are unable to answer that question for you. But perhaps it will always be a rare conjunction of circumstances that brings together an education minister who heeds expert advice and who is aware of the existence of the right moment for releasing the potential for change when that expert advice is transformed into action.

The argument for the abolition of external examinations was political as well as theoretical. It was about control over the senior curriculum by the University of Queensland. In the words of acting Vice-Chancellor Davies (University of Queensland, 1978):

I believe that the motive for change from the Senior Examination to school-based assessment was the need to remove the constraints it placed on the secondary school curriculum.

[...Although it was introduced too quickly...] teachers are beginning to accept greater responsibility with enthusiasm and there is now a much greater degree of professionalism among teachers than before.

Forty years on from the infamous 1967 Physics paper and discussions about school-based assessment again refer to the notion of control, specifically control over the setting and marking of high-stakes assessments. External examinations and school-based assessments are obviously not the same in terms of their loci of control for setting and marking. External examinations, like standardised tests, have an external source (a central agency) controlling both setting and marking. School-based assessments have an internal source controlling both setting and marking (subject teachers). But the argument now tends to be framed in terms of the advantages and disadvantages of using teachers' judgments in high-stakes assessment programs.

The same theoretical discussions occur now as in 1968. They are essentially about validity and reliability. Wiggins (1991) captures this dilemma when he states:

To ask about validity is to ask if the task represents the real thing we want to assess. Does it really represent the student's abilities, traits, capacity for long-term work? For example, the SAT is valid because it statistically correlates with

later success in college. But does it really represent the things the student can be good at, or just one thing? Reliability is another question. Would the student get the same score if she took the test again and gave the same performance? Or would different people score it differently? Standardised tests are reliable by design but we question their validity. Exhibitions, on the other hand, are valid but not necessarily reliable. How do we protect students from capricious, biased judgments?

The Queensland senior assessment system is an operational model that puts teacher judgment at its centre while grappling with the issues of validity, reliability and equity. What happened to such a system over 40 years is of crucial importance in continuing the discussion precipitated by Radford and highlighted by Wiggins. Before this however, we build into the frame consideration of the national agenda.

The national agenda: Australian government

Under the Australian Constitution, education is a state responsibility, and there are eight independent jurisdictions (six states and two territories). All issue senior secondary certificates at the end of Year 12, the final year of schooling. There are significant differences between jurisdictions in arrangements for curriculum, assessment and certification. There are differences in procedures for ensuring comparability of standards in reported results and in procedures for combining results from different subjects in compiling tertiary entrance ranks. These and many other differences are grounded in the history of the states and territories and their education systems, and in the different sets of compromises that have had to be struck by curriculum and assessment agencies with their respective stakeholders over the years.

The Australian Government recently investigated the introduction of a single Australian Certificate of Education in pursuit of greater consistency in senior secondary arrangements for curriculum, assessment and certification, more comparable student results across jurisdictions, and clearer and more consistent standards of student achievement (Masters, Forster, Matters, & Tognolini, 2006).

The Government then commissioned a study (Matters & Masters, 2007) that responded to these questions about five subjects³ in the final year of schooling nationwide:

- Curriculum content: What is common?
- Curriculum content: What is essential?
- Achievement standards: Are they comparable?

The study established significant consistency in what is assessed – and this is a prerequisite for meaningful comparisons; however, different jurisdictions use different methods of assessment (e.g. external examinations, teacher-devised assessment instruments and projects), raising the question as to whether achievement standards can be compared across jurisdictions, or whether the existence of different assessment methods confounds comparison.

Two overarching recommendations came out of this study (the so-called ‘CAS study’):

1. Identify, for each of some nominated senior school subjects, a curriculum ‘core’ that clearly specifies what all students in Australia taking that subject are expected to learn, regardless of where they live in Australia.
2. Develop a set of achievement standards as a nationally consistent description of how well students are expected to learn the core in each subject.

The report pointed out that the call for greater consistency, increased comparability, and clearly stated achievement standards did not necessarily imply the need for a national curriculum or common national subject examinations. The report did note, however, that

³ English (including Literature), Mathematics, Chemistry, Physics and Australian History

the achievement of greater consistency, comparability and clarity in these areas is inevitably more difficult if the underlying certificates, curriculum and assessment programs are independently developed and managed (Matters & Masters, 2007: 2).

In February 2008 the Government announced the establishment of the National Curriculum Board (www.ncb.org.au). This was followed on 2 May 2008, by a media release from the NCB, which stated that 'the Board had moved immediately on its first priority ... to develop a national curriculum in English, mathematics, the sciences and history', that it would be a 'collaborative process, with Australia working together to produce a world-class curriculum', and that 'what would be considered first would be the best relationships between content and outcomes, between common national content and regional variation and between curriculum specification and school and teacher discretion'. The title of the media release was 'National curriculum journey begins'. The title was very apt since one certain thing about the Australian psyche is that Australians define themselves according to their state of origin before they define themselves nationally.

The slippage from the notion of 'core' in the CAS study and the notion of a national curriculum is dramatic. The term *core* was deliberately used in the CAS study as a noun or, sometimes, noun as adjective. The report argued for national *standards* in the core or *essence* of selected disciplines but, beyond that, no journeys in pursuit of consensus about what knowledge is of most worth across Australia at the beginning of the 21st century. That question remains unanswered since Herbert Spencer asked it in 1884. And now we have added complications (e.g. formal curriculum content versus generic skills; states versus the commonwealth; empiricism versus consensus building; statements of curriculum intent versus evidence of curriculum taught and learnt and assessed and so on).

The story to date: Sign posts

This section draws on Smith (1995)⁴ to provide a more extended historical sketch of the transformations that have occurred in Queensland's assessment policy and practice from 1876 to the present, expressed in terms of four eras.

Era 1 (1876–1969): The reign of external examinations

Public examinations were first held for Queensland secondary school students in 1876 and persisted well into the second half of the 20th century. The University of Sydney was responsible for the setting and marking of these examinations until 1912 when the University of Queensland took over this role (for Queensland students) after coming into existence as Queensland's first university. Results were published for 'passing' students as A, B or C until 1967 when results were reported on a scale of 7–1, covering all students who sat the examination. This change in nomenclature for reporting and certificating confounded the infamous performance of the 1967 Physics subject-group, mentioned earlier in this paper. The Physics professor who set the paper had examined topics outside the syllabus.

The main challenges for teachers in this era were to ensure that they covered the intended curriculum and that they prepared their students for the ritual of the examination and for the content of the examination.

Over time, the examinations had inevitable and strong backwash effects on the curriculum and classroom teaching, learning and assessment. Routinely the teaching year was staged to build student knowledge of the type required for display in the examination, with rehearsal for the type of questions and for managing time restrictions under examination conditions. In large part, the examination items focused on student control of content knowledge, the time-limited examination genre not permitting opportunities for extended problem-solving or evaluative thinking that require more

⁴ Now publishing as Wyatt-Smith

time and access to material resources. In retrospect, it is fair to state that, irrespective of the quality of the examination in any given year, the scope of the knowledge, skills and capabilities assessed was very narrow, relative to that routinely taught and assessed in accordance with extant syllabus materials in the State. Also, the examinations worked to define the roles of the teacher and student as both pitting themselves against the demands of the examinations, with past papers providing rehearsal opportunities. Further, the grading of student work relied on numeric scoring tied to a reporting framework using letter grades, in the absence of any sense of quality represented in standards stated as verbal descriptors. In the latter phase of the public examination system in Queensland, student results in the form of letter-grades were published in newspapers, the grade appearing with the student name and school attended. In part as a legacy of this era, there remains in the community, and to some extent in the media, residual understandings that numeric scores captured as percentages have some absolute, or at least intrinsic, meaning.

Era 2 (1970–1978): From external examinations to school-based assessment

The first major shift – and possibly the most radical – occurred in the early 1970s with a shift from external examinations to school-based assessment. This shift, which was precipitated by ‘the increasing disenchantment with the public examination system’ (Sadler, 1993: 3), occurred when a key recommendation in Radford’s report to government was accepted: The movement to fully school-based assessment became known as the Radford Scheme, which, as Smith⁵ (1995: 11) explains:

...represented a radical change which was without precedent in Australia, [and internationally] and pioneered norm-referenced school-based assessment using teacher-made tests. In essence, the scheme involved a significant devolution of authority for assessment to the classroom teacher, the school and review panels, and a shift in emphasis from terminal (final) to continuous (ongoing) assessment. No longer was it the teachers’ responsibility to prepare students as candidates for external, centrally controlled examinations. Rather, for the first time in the history of secondary education in Australia, teachers were required to document the main aspects of a course of study; to develop and implement a range of test instruments including assignments and examinations; and to report on student achievement [using a norm-referenced approach].

All high-stakes assessment regimes are concerned about the reliability of results produced and so in this era a process of social moderation was introduced to ensure that grades assigned between schools across the state were comparable. At this stage of evolution of the system, the technique of social moderation that was practised was validation of teachers’ judgments by their peers at teacher meetings. There was an unexpected benefit of this process, retrospectively described by Sadler (1993: 3) as ‘the greatest influence on the professional development of secondary teachers in Queensland’s history’. The opportunity for teachers to come together to discuss and observe the work occurring in other schools proved highly beneficial.

The challenges in this era were for teachers to design good assessment programs and assessment instruments (the adjective ‘good’ being used deliberately to connote all desirable properties) and for teachers to be able to critique the work (assessment instruments and decisions about proposed distributions of grades) of teachers from other schools.

Era 3 (1979–1985): From norm-referencing to criteria-based assessment

The second major shift was prompted by several concerns with the norm-referenced approach to assigning grades (or levels of achievement). Two research reports (Campbell et al., 1975; Fairbairn et al., 1976) concluded that the ‘norm-based awarding of grades contributed to unhealthy competition and even animosity among students’ with a

concomitant 'erosion in teacher-student relationships' (Smith, 1995: 12). Additionally, tests and examinations increased in frequency rather than decreased as was expected. An expert panel chaired by Professor Ted Scott of James Cook University was appointed to review the two research reports and provide advice on future policy and practice. Recommendations in 'A Review of School-Based Assessment in Queensland Secondary Schools' (acronym ROSBA) were implemented in three successive phases from 1981 to 1986. Most significant was the change from norm-referenced assessment to criteria-based assessment (a completely new approach to awarding grades). Elements of the Radford Scheme such as continuous assessment and teacher responsibility for planning and implementing programs of work and reporting on student remained as elements of ROSBA. In retrospect, the term 'ROSBA' was an unfortunate way to describe what was in reality standards-based assessment, albeit where the assessments were carried out by teachers as expert judges.

ROSBA represented a significant change in mindset, from heavy reliance on direct comparisons between students to the application of criteria and standards as the yardstick for awarding grades.

An underlying premise of the system was that student performance can be improved if teachers make available the criteria to be used in judging the quality of student performance. In practice, ROSBA required that teachers prescribe and publish detailed criteria prior to student commencing an assessable task (Smith, 1999: 13-14).

The challenges in this era were for teachers to pioneer a new approach while inexperienced in the use of criteria and standards and to a certain degree confused about the difference between 'criterion' and 'standard', and for the Board to support teachers in coping with the ambiguity of the time (a nonetheless exciting time). There were no models to copy even if the Board had wanted to be derivative. Hence the Queensland senior assessment system can be portrayed as 'not so much underpinned by theory but as a theory-building exercise in itself' (Matters, 2006a: 6). The theoretical foundations of the criteria/standards approach were laid when the Board obtained government funding for an Assessment Unit which produced more than 20 discussion papers, primarily for a teacher audience. Whether teachers read the papers and, if they did, what impact their understanding of the theoretical issues had on policy and practice are unclear.

An author of this IAEA paper has her own original set of 21 discussion papers in retro pastel blue shiny covers. Some of the work inside those covers was definitely before its time, and many of the issues explored by Sadler, McMeniman, Beasley and Findlay (1986) remain current today (2008); for example, the following explanation by Smith (1995: 18):

A key premise underlying this organisational feature of the system is the proposition that formative and summative assessments are not mutually exclusive but complementary approaches to providing a reliable indication of student achievement (McMeniman, 1986b). A related premise is that classroom teachers are in the ideal situation to monitor their students' learning, and also to provide informed judgments and reports on student achievement.

Era 4 (First phase 1986-2003): Movement from criteria-based to standards-based

Since 1986, a school-based approach to assessment, known as criteria-based assessment, has been operating in all Queensland secondary schools. A key feature of this model of assessment is that teachers judge the quality of student work (either a single piece or a representative sample) by comparing it with (matching it to) explicit criteria and standards.

An underlying premise of this approach, as mentioned above, is that student performance can be improved if the teachers define and make available to students the criteria that will be used in assessing their work. This means that, in principle, students no longer need to guess at what teachers will value in student work (responses, performances, artefacts or whatever the output of the assessment task). A related premise

is that students will see that their work is not judged against teachers' implicit standards but, rather, against the elaborated standards for the assessment criteria. Ultimately, reliability and the credibility of teachers' judgments are enhanced.

This (fourth) era was characterised by developments in the conceptualisation of school-based assessment that reinforced the relationship between criteria and standards, taking stated criteria to its centre and, in turn, defined standards, written as verbal descriptors of quality.

Pitman and O'Brien (1999) insist that standards do not exist as words on a page. The existence of standards statements is but one element of three essential elements; the other two are the instantiation of standards in student work and the agreement among teacher-assessors (expert judges) that a certain piece of work (single or collection) meets a certain standard.

Although the collection of discussion papers from the Assessment Unit went some way to providing a theoretical framework, a comprehensive and fully articulated version of the underlying theory of criteria- and standards-based assessment in Queensland is not available some 27 years after the implementation of ROSBA began. This situation can be accounted for, in part, because the Assessment Unit was disbanded in the late eighties as a result of funding cuts. Since then, there have been no significant developments in the underlying theory of the system, from either a curriculum perspective or its assessment dimension, apart from how the developing system documented itself and provided insights (see Appendix 2).

One of the challenges in this era was firming up on the link between assessment criteria and standards, which was accomplished by the end of the era, leading to a concentration on stated standards and reporting.

During this period, social moderation (for the purpose of validating teacher judgments) occurred through panel review rather than through teacher meetings. The philosophy of peer review remained but the technique changed for pragmatic reasons (not the least of which is the size and diversity of the state of Queensland). For a detailed description of social moderation and statistical moderation see Matters (2006c), Maxwell (2007), Pitman, O'Brien and McCollow (2002). Details of moderation in action, policy and procedures for schools and logistics at the central agency appear in various moderation handbooks and guidelines from QSA since 2002 (or Board of Senior Secondary School Studies before that).

Era 4 (Second phase 2004-2008): Strengthening spotlight on standards and reporting

Towards the end of the fourth era there was a discernible move in research, policy and, to a lesser extent, practice, towards having stated standards and issues of quality at centre stage. One catalyst for this move was *The Review of Tertiary Entrance in Queensland* (Viviani, 1990). It called for an evidentiary base reflective of the education system's efforts to subject itself to scrutiny and to provide data useful for evaluative and improvement purposes. This resulted in the formation of the Board's Research, Evaluation & Development Section. There were two other catalysts of note: the New Basics research program (Department of Education and the Arts, 2004) and work done under the banner of the Assessment and Reporting Framework Implementation Committee (Education Queensland, 2004). While these two initiatives were radically different in nature, purpose and scope, common to them was a commitment to install a system that aligned curriculum, pedagogy, assessment and reporting, with the strong focus on teacher knowledge of task demands and stated standards. Indeed, words and actions that prioritised the alignment mentioned above already existed in Queensland at the time of the federal government's decision regarding a common reporting framework and the state government's launch of the Queensland Curriculum Assessment Reporting (QCAR) Framework (QSA, 2008). While the latter is outside the scope of this paper, the QCAR

initiative is an instance of the prominence of standards and reporting in educational policy directions at state and federal levels.

Looking across the eras

In looking back we see the evolution of an externally moderated model of school-based assessment in senior schooling that has a complex set of quality assurance mechanisms. These include primary sources for evidence upon which to base decisions about achievement, teachers making those judgments about achievement, peer review (teachers verifying the judgments of other teachers), explicit standards, and demonstrated accountability. And all of this has been achieved through intellectual commitment and successful change management at all levels of the system. (Readers interested in a more detailed discussion of the characteristics of the system and key decision points in its development are advised to see Attachment 1.)

As an instance of the demonstrated high regard that the system has previously achieved, consider the use of the terms 'revolutionary', 'remarkable' and 'excitement' in the segment below from a speech by Carol Myford from ETS in 1999:

When people ask me who is on the cutting edge, my first response is, 'look down under' ... On re-reading a 1985 account of Queensland's externally moderated school-based assessment, I remember thinking how truly revolutionary it was in scope. Upon my second reading [2000] and taking into consideration the political realities of the late 1990s ... I find it even more remarkable. My reaction to this program has moved up at least two notches on the excitement scale.

Gunn (2007), however, in reporting on her search for scholarly work about aligning large-scale formative and summative assessment, supports Matters' observation (2006a) that many of the papers cited in the education literature are those written by Queensland academics (e.g. McMeniman, Sadler, Beasley) and key figures in the implementation of criteria-based assessment (e.g. Pitman, O'Brien, Dudley). Beyond this, the Queensland model has been used as an exemplar in international scholarly articles (e.g. Elwood, 2006; Gipps, 1996; Gipps & Stobart, 2003; Harlen, 2004a, 2004b; Myford, 1999; Shavelson, Black, Wiliam, & Coffey, 2004; Strachan, 2002).

Shavelson et al. (2004), for example, discuss cases in three jurisdictions (United Kingdom, California, and Queensland) that have attempted to align large-scale summative and formative assessment with varying degrees of success. For California, the devil was found to be in the detail of implementation (and politics). For the UK, a struggle for power between competing ideologies caused dismemberment of what had been recommended (Shavelson et al., 2004: 18). For Queensland, struggling with a 'theory-building exercise' and teachers unprepared for radical change, the recommendations for change were able to be implemented and successive governments allowed the model to evolve, a process which continues today. While the system's claim to fame has been its success in practice, this has not been theorised as a real-world manifestation of standards-based assessment.

Looking sideways: Challenges on four fronts

The existence of an internal assessment regime is at a time of critical review, given the following challenges:

1. The federal⁶ government's directions in national curriculum and standards.

⁶ Current federal education priorities include but are not restricted to national curriculum as an improvement mechanism, together with common assessment and examinations. Literacy and numeracy continue to figure prominently in relation to schooling performance, workforce productivity, and national long-term prosperity.

2. The consequences of the ways in which an organic system develops and is subject to change in practice.
3. The need for systematic induction of the profession (practitioners and bureaucrats) into the key tenets of the Queensland assessment model.
4. The need for provision of systematic documentation of the change processes that Queensland undertook in installing and maintaining the model.

The national agenda and the national curriculum board are certainly big news but they are not the whole story. Even without the national agenda, it would be timely for the system to reflect upon its current practice, especially given the policy directions in the State towards a single assessment framework across the years of schooling. For the Draft P-12 Assessment Policy readers are advised to see: <http://www.qsa.qld.edu.au/assessment/3111.html>

Further attention needs to be given to strategies for enabling deep understanding of fundamental concepts that underpin best-practice principles in standards-based assessment in Queensland or, indeed, of the pros and cons of internal and external assessment, originally presented in Radford (1970). Limited understanding of internal and external assessment can be partly explained by intergenerational change, though accounting for the unevenness of understanding about the principles of standards-based assessment is more difficult.

Drawing on our experience as outlined earlier, and without providing more detail than is necessary, we point to challenges that have developed in several crucial areas:

- In the area of subject advisory committees, factional debates can occur instead of scholarly exchange.
- In the area of panel advice (moderation) – de facto rules ('rules of thumb') can grow up over time and prevail over recurrent consideration and reflective understanding.
- In the area of the format for a standards schema for assigning exit grades, two issues combine namely the language used to capture expectations of quality and information provided about expected judgment practices.

Of course, factional debates, de facto rules, and 'wordy' assessment rubrics are not confined to Queensland. And, going back further, a scrutiny of other systems would no doubt bring to light less-than-perfect understandings of their fundamental concepts. But other systems are more orthodox in the sense that they are less likely to make possible a wide variation in understandings of such concepts in the profession.

Looking forwards: Realising potentials in the next era (beyond 2008)

The education landscape in Queensland (as in the rest of Australia) is extremely fluid. Part of this is related to: short political cycles; to the change in government at the national level, reported earlier, and to shifts in education policy direction (e.g., the shift away from outcomes to essential or core curriculum in several states). If we acknowledge the magnitude of the rate of change in the education landscape possible in terms of power, influence, expertise and accountability, then we must acknowledge that the chance of there being continuity and coherence in assessment policy and practice is limited. Arguably then, if the background changes continuously, the foreground needs to be more predictable, otherwise the effects of disequilibrium on schools and students can become overwhelming.

We propose that the Queensland senior system is well placed to present and receive messages about standards (as external referents of quality), criteria (and the role they play in specifying desirable features of performance), quality assurance mechanisms (checking demands and rigour of teacher-devised assessment instruments and programs relative to curriculum intent and stated standards), and the teacher's role (especially in enhancing students' evaluative experience through their use of pre-set standards to

inform improvement efforts). The possibility of fusing internal assessment and standardised examinations may not be welcomed, but may be part of a direction at national level. We may have reached a stage in the decision tree (see Figure 2) where a shift of direction is to be considered. Our key point is that, ultimately, the alternatives are not external examinations versus internal (school-based) assessment. The alternatives are good assessment (setting and marking) versus bad assessment instruments (setting and marking). And this ultimatum applies to any existing system, be its assessment regime external, internal or a combination of the two. But if a decision is to be made, it needs to be based on a sound knowledge of the differences between internal and external assessment. There will be a need for serious debate about these differences if the Queensland model is to be more fully understood nationally, and if relative merits of different regimes are to be examined seriously.

Essentially, the Queensland model requires that the partnership between schools and the curriculum–assessment agency be central; that assessment be unashamedly situated at the epicentre of teachers’ curriculum planning, and that achievement standards be the unifying device for the assessment system. For achievement standards to work in this way there must be high-quality assessment instruments (for bringing forth evidence of learning) and high-quality judgments of that evidence (for ensuring comparability).

Conclusions

To return to where we began with this brief historical sketch – we argued that history is able to provide a perspective on the present, improve understanding of current situations, and raise questions that might not otherwise be evident. So what have we learnt, what questions do we need to ask and, importantly, what don’t we want to repeat.

We learnt that we cannot assume the relevance of teachers’ prior experience when embarking on a radically different approach to assessment or for that matter any other educational reform. Implementation ideally should follow a deep professional knowledge and understanding of the system to be implemented by all concerned, particularly teachers. After this, the focus shifts from details of implementation to concern for the preparedness of those who must implement the new system. It also calls into question the wisdom of embarking on reform as a theory-building exercise rather than reform taking place within an established theoretical framework.

Assessment reform involving such major shifts inevitably raises questions on issues about the management of change processes and the impact of these shifts over time. Matters (2006a) alludes to some of these issues when, over 20 years after the implementation of non-norm school-based assessment, there continue to be countless interpretations of criteria/standards and much confusion in notions of assessment, the nature of criterion, assessment instruments, and assessment technique. It would therefore seem that a common understanding of the language of assessment used in the Queensland system is yet to be established, some two decades after the language was adopted. As the State of Queensland was one of the first nationally and internationally to make significant shifts to school-based, followed by non-norm referenced, assessment questions regarding the impact of such radical change in terms of teaching, pedagogy, learning and classroom assessment practices are of particular interest – questions still to be comprehensively researched.

At the beginning of this paper it was suggested that the impetus for these radical changes in assessment practice was disenchantment with an external examination system which was considered too narrow and as having undesirable consequences (Sadler, 1993; Shavelson, Black, Wiliam & Coffey, 2004). A recent refrain at both national and international levels for public institutions, including schools, to be held accountable has led to a push for the return to external accountability systems. This has the potential in Queensland to re-establish the oft-reported tension between information to improve teaching and learning, and information to inform the public of education quality. As Shavelson et al. (2004: 35) warn:

While polls show widespread support for the noble democratic concept of accountability, accountability can and does fall short in practice. When the stakes are high, as they are now in education accountability systems, and when the interpretations of large-scale assessment scores with ambiguous or narrow meaning are treated in league tables and funding decisions as unambiguous, and when single scores are generalised beyond justification as true characterisations of individuals and systems, the potential for mischief is enormous.

Queensland has journeyed some distance along the challenging assessment road over the last 40 years and it would be ironic if the accountability demand with a focus on external assessment, ostensibly to improve education provision, became the biggest impediment to achieving that improvement. The possibility of fusing internal assessment and standardised examinations should not be rejected out of hand.

Figure 1 is a diagrammatic representation of the story of Queensland's four assessment eras for the time-span 1912 to 2008.

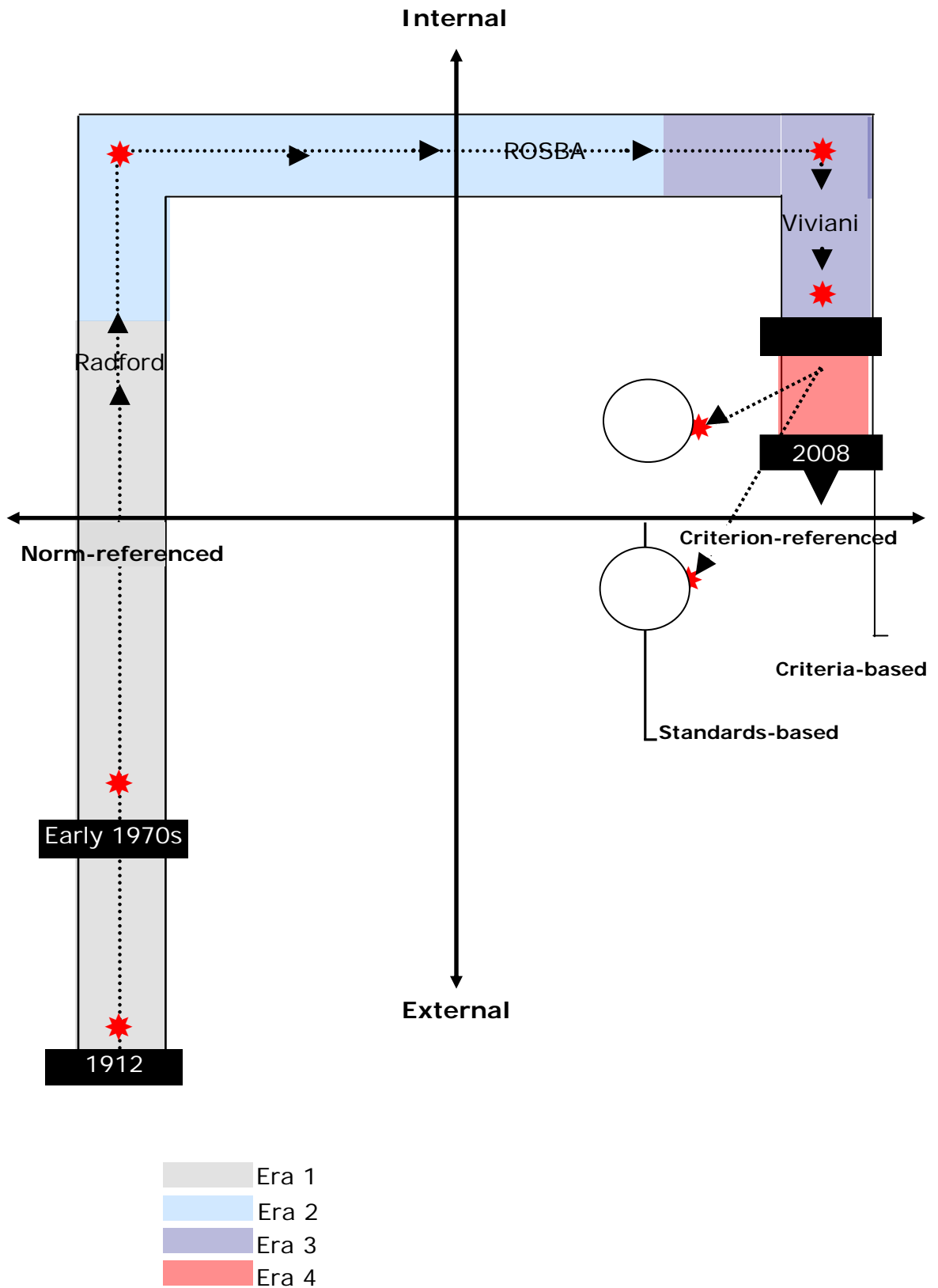


Figure 1: Historic overview of the evolving system of school-based assessment in Queensland, Australia (based on Pitman, 2002)

Attachment 1

Alternatives in assessment arrangements: Reviewing decision points

Assessment regime

Assessment arrangements can be classified as internal or external, and assessment regimes use internal, external, or a combination of the two. In this paper, external assessment is taken to refer to external examinations. These are subject-specific examinations set by a body external to the school, as exemplified by the Cambridge A-levels in England. Such examinations are devised to assess student achievement in a particular subject, whether by objective-type or conventional written, oral or practical questions. All the questions refer to a syllabus that has been defined by a group of educators (including teachers and/or examiners) and signed off by the government of the day.

External examinations are point-in-time, time limited and standardised (that is, common paper administered to all candidates under commonly applied conditions, and marked according to a common marking scheme). Internal (school-based) assessment is devised, constructed and implemented by schools, sometimes based on an official syllabus and accredited work program, sometimes not.

Different modes of assessment (e.g. multiple-choice, constructed response, extended writing) dominate in different assessment regimes. A wide variety of assessment instruments can be used as tools or devices or constructed situations for bringing forth evidence about student achievement. (e.g. written assignment, oral presentation, performance, demonstration of mastery, practical work, field work, test, examination, project, thesis, and viva voce). And there is a range of assessment techniques that can be associated with each type of assessment instrument: Evidence about student achievement can be gathered through observation, consultation, or analysis of student work (holistically or analytically).

For students and schools particularly, much is at stake when participating in assessment regimes, whether external or internal or a combination of both. The term 'high-stakes assessment' is used in two senses: One is associated with the important consequences for the student (e.g. being promoted or receiving a certificate); the other is associated with consequences for instruction quality (e.g. school rankings). The definition of high-stakes applies to senior assessment in Queensland in both senses of the word.

Moderation

All high-stakes assessment regimes are concerned about the reliability of results produced. The response of the current Queensland model to the reliability challenge traditionally associated with school-based assessment is social moderation, in which contextualised teachers' judgments about the quality of student work are verified by the teachers' peers. It is based on the premise that the teacher-assessors are expert judges.

Validating teacher judgments in internal assessment is one situation in which comparability is desired. Another is putting results onto a common scale. There are at least five different approaches to linking results from different assessments (Linn, 1993: 93). The approach taken depends on the purpose being served. Two⁷ familiar approaches are social moderation and statistical moderation.

Examples can be found of matches (✓) between the abovementioned purposes and approaches or forms (see Table 1). Table 2 provides familiar techniques associated with each of the forms. The bold-face text locates the purpose, form, and technique of moderation currently employed. It should be noted that the Queensland Core Skills (QCS) Test is not discussed in this paper. A common

⁷ Linn (1993) mentions three others: Equating, Calibration and Prediction.

misconception is that the QCS Test is used to validate teachers' judgments. It is not. Group-parameter from the QCS Test are used in the process for combining results from different subjects in different schools to compile a statewide rank order list for tertiary entrance procedures (see Matters, 2006c).

Table 1: Purposes and forms of moderation

Purpose	Form	
	Social moderation	Statistical moderation
Validation	✓	✓
Scaling	×	✓

Table 2: Validation techniques in social moderation

Form	Technique
Social	Teacher meetings
	Review panels
Statistical	Visitation by 'moderators'
	Using external examinations
	Using an aptitude test

Assigning grades

It could be argued (Matters, 2006a) that the increasing confidence about school-based assessment over the years has been accompanied by a decreasing level of knowledge about other ways of doing things, including alternative ways of assigning grades. At the same time, there has been a significant influx of teachers from other parts of Australia and from abroad. Many of these teachers have not been part of a culture of criteria/standards-based assessment and the majority of them would have learnt about the operationalisation of externally moderated school-based assessment from their (new) colleagues. Sometimes much is lost in translation.

Sadler (2000) states that the term 'criteria' is seen to have sacred status in Queensland and the 'lack of knowledge about alternative ways of assigning grades has suppressed intelligent discussion and practical progress'. While his criticism was shared with educators in higher education and presumably refers more to them than to educators in the senior schooling sector, the point is pertinent to this discussion about alternatives in assessment arrangements.

It is the case that, more than 20 years after the introduction of criteria-based assessment into Queensland, teachers who are new to the profession and/or new to the State are obtaining knowledge of criteria-based assessment through translations and re-interpretations of the topic from their colleagues in staffrooms around Queensland. Such information giving might explain the lack of knowledge of the alternatives (whether these be good or bad) that could be used in the process of assigning grades, and the muddiness that now seems to surround establishing and maintaining standards.

The section below summarises Sadler's (2000) four alternatives for assigning grades.

Setting numerical boundaries for grades

The most commonly used grading scheme throughout the world is to award grades according to pre-determined numerical ranges such as A+ (90–100), A (80–89), B (65–79), and C (50–64). The criteria are the grade boundaries. Students are

not graded against one another, but according to whether they reach pre-set standards as specified by the mark ranges. This is what underpins many teachers' mark books and spreadsheets. In some place at some time, the grade boundaries were set as policy (e.g. results in the Senior Public Examinations until 1966).

Although numerical boundaries (or 'cut scores') no longer dominate in subject-specific assessment (if used at all), it is the case that, in the Queensland Core Skills (QCS) Test, cut scores are applied alongside standards referencing.

Combination rules on pre-determined criteria

A grading scheme that has become increasingly common throughout the world is to communicate to students at the beginning of a course what the criteria are for assessment (e.g., for exit assessment and task-specific assessment). Students are informed about what the assessment program will be, its components, and how results will be combined. In the case of overall or exit assessment, the composition rule, which is formulated by syllabus writers, states how the results are to be combined and grades assigned. The criteria, which sometimes incorporate mandatory minima on certain components, are the composition rules. Nominating the criteria does not tell students about standards.

Characterising a student's global achievement in a semester/year/2-year course in terms of exit criteria

In this way of assigning grades, the criteria are the desired characteristics of a student's global achievement in a course. They connect with the idea of generic skills. The description that accompanies each grade is given as a guideline to assist comparability across the State, but these descriptions have to be interpreted within the context of the subject's delivery in a particular school.

Properties or characteristics or qualities of a particular piece of work

These characteristics of a particular piece of work (i.e. one instance of assessment) are different in both scope and kind from the characteristics of a student's global achievement in a course of study. A general flavour in the scoring rubrics might be appropriate for application to tasks or dissertation but topic- or task-specific criteria would also have to be stipulated.

The current model characterises a student's global achievement in a 2-year course of study in terms of exit criteria and judges it against standards (i.e. an elaboration of each of the criteria).

Criteria-based or standards-based

Teacher-assessors judge the quality of student performance on multiple criteria with reference to pre-stated standards. Are we talking about criteria-based assessment or standards-based assessment?

Queensland's senior assessment system has been called *criteria-based* because it focuses on the specific nature of a student's actual achievements in relation to specific criteria (rather than to an established norm or relative to other students). But the primary focus is on standards rather than on criteria, although standards pre-suppose criteria. This leads to an appreciation of why the system can be called standards-based, representing a perturbation at the end of the fourth era (see Figure 2 later). In their purest form, standards are descriptions or other specifications of performance levels that are free from any reference to the performance to the typical student, and the proportion of students expected to achieve at a given level, or the particular age, or stage of schooling at which a certain level of performance is thought to be acceptable. Because the defining feature is the notion of matching student work to pre-determined standards, the Queensland system is now taken to be *standards-based*. This shift in emphasis subtle; the introduction of the new label went unheralded.

The philosophy, policies and procedures in the senior system now refer to a system that is actually standards-based (compared with standards-referenced or criteria-based). To complicate matters, Tognolini (2005) defines standards referencing as the process of giving meaning to marks assigned to student work by referencing the image of the work to pre-determined standards of performance. This definition is most suitable for New South Wales with its emphasis on marks. Both definitions, however, share the notion of matching student work to pre-determined standards. The distinction in practice is one of sequence.

The problems associated with the ambiguity of the term 'criteria' explain some of the current jumbled attitudes to the implementation of criteria-based assessment in other jurisdictions. Sadler (2000) provides two reasons for the problematic nature of the use of the term criterion: one, the term 'criterion' is often used when 'standard' is meant; and, two, the term 'criteria' has multiple meanings.

In everyday conversations, the words criteria and standards are often used interchangeably even though they are defined differently in the dictionary. In assessment conversations, the distinction between the terms criteria and standards is one that breaks the process of teacher judgment into two stages. First, the criteria have to be identified; then the standards on the various criteria have to be specified.

Sadler's (1987: 194) definitions follow.

Criterion *n.* (*pl.* criteria): A distinguishing property or characteristic of any thing, by which its quality can be judged or estimated, or by which a decision or classification may be made. Derived from the Greek *kriterion*, a means for judging.

Standard *n.*: A definite level of excellence or attainment, or a definite degree of any quality viewed as a prescribed object of endeavour or as the recognised measure of what is adequate for some purpose, so established by authority, custom, or consensus. Derived from the Roman *estendre*, to extend.

For the student, the highest standard is a goal to aim for. For the teacher, 'standards are the referents that underlie judgments about success or level of merit in a performance' (see Maxwell, 2001). In evaluating the quality of student performance, the teacher judges which one of several designated standards best represents the characteristics of a student's performance; that is, what label to attach to the performance or what category (such as A-E) in which to place it. Each identified criterion has associated standards for A-grade student work (note it is the work that is A-grade not the student). B-grade student work would use the same criteria but the standards would be lower. A lower standard might be described in terms of quality, quantity, sophistication, length etc., or permutations thereof. This is the art of setting and describing standards.

The combination of the standards on the (say three) criteria might be called collectively the 'overall standard'. In Queensland, the combination gives an exit level of achievement, and this is coded as VLA, LA, SA, HA, VHA. Combining judgments is not the same as combining scores.

Combining scores involves weighting the components of an assessment program according to standard deviations of the scores on those components before combining the results to produce aggregate scores and ultimately grades. Different assessment item types and formats (such as essays and objective tests) commonly produce different levels of differentiation among students. Objective tests and problem solutions tend to produce greater variability than essays. Weighting according to standard deviations ensures that the results from the separate components contribute appropriately to differentiation in the aggregate scores. This process, which has a normative element in it, does not require formal prior

definition of standards or precise estimates of how students will perform on assessment tasks. Clear expectations as to the exact nature of the product are not required because there is no necessity to specify it in detail or model it in advance.

Combining judgments on separate criteria involves trading off

Trading off in the assessment context is the process of making on-balance judgments about the standard of student work. It ensures that good performance on one criterion can compensate for poorer performance on another; and that performances on several criteria contribute to the grade assigned to student work in a manner reflective of their hierarchical positions (if, of course there is a hierarchy); that is, of the ranking of the criteria in a way that is indicative of their relative contributions to the award of grades.

Whether or not this should be called weighting is another matter. If it is a case of weighting, then the rules of combination require statistical methods. Either way, the notion of trading off or making on-balance judgments is crucial to deciding a student's exit level of achievement.

The primary reason for using a criteria/standards matrix (the way exit levels of achievement are presented in syllabuses) is to ensure validity. For each individual student, teacher-assessors must make judgments on the same basis so that there is an explicit match between this basis and the standard stated in the syllabus. If only one piece of information per student per subject (i.e. one grade expressed as a level of achievement) is captured for certification then the question must be asked: To what extent will trade-offs be allowed and how will trade-offs between differential criteria be facilitated? There are at least five answers to this question, not detailed here (see Matters, 2006a).

In Queensland, the terms grade specifications and standards descriptors seem to be used interchangeably. The difference, however, is that grade specifications state what the various levels of achievement (grades) are to consist of whereas standards descriptors state what student work for the award of particular levels of achievement (standards) is to be like. A standards descriptor, being a statement or list of statements that succinctly conveys the required quality of, or features in, student work in order for it to be awarded the corresponding grade, could operate at the domain level or the task level or the examination level or on exit from a course of study. In summary, the process of grading occurs under the following conditions.

- Teacher-assessors use dimensions (criteria) provided to interrogate the evidence and make judgments about the quality of student work.
- There are various ways to satisfy a standard in terms of the evidence tendered.
- Student work that is of equivalent standard should be awarded the same grade.

Figure 2 traces the route that Queensland took through these possibilities. That route can perhaps best be understood in relation to the roads not taken. Decisions were made at various points: in terms of regime (e.g. was it to be external or internal or something else?), assessment process (e.g. was it to be normative or criterion-referenced or something else?), method of grading (e.g. was it to be through combination rules on pre-determined criteria or numerical cut-offs or something else?) and so on. At each junction, one road was taken and others were not (necessarily). The roads taken and not taken can be represented as a decision tree, as in Figure 2. The red lines in the decision tree are those that the system or policy makers took in making their choices. Accordingly, Figure 2 provides the record of Queensland's assessment inheritance.

The architects of the system, the (then) Board of Secondary School Studies and its committees, had to draw on the limited literature and experience elsewhere, and then create their own operating system.

Where the current Queensland senior system has arrived along this route can be classified as externally moderated school-based standards-based assessment in a high-stakes environment. Assessments are devised and marked by teachers, judgments are validated through the panel model of social moderation, and grading is based on the application of a standards schema.

Defining Elements of the Current Approach to Externally Moderated School-Based Assessment

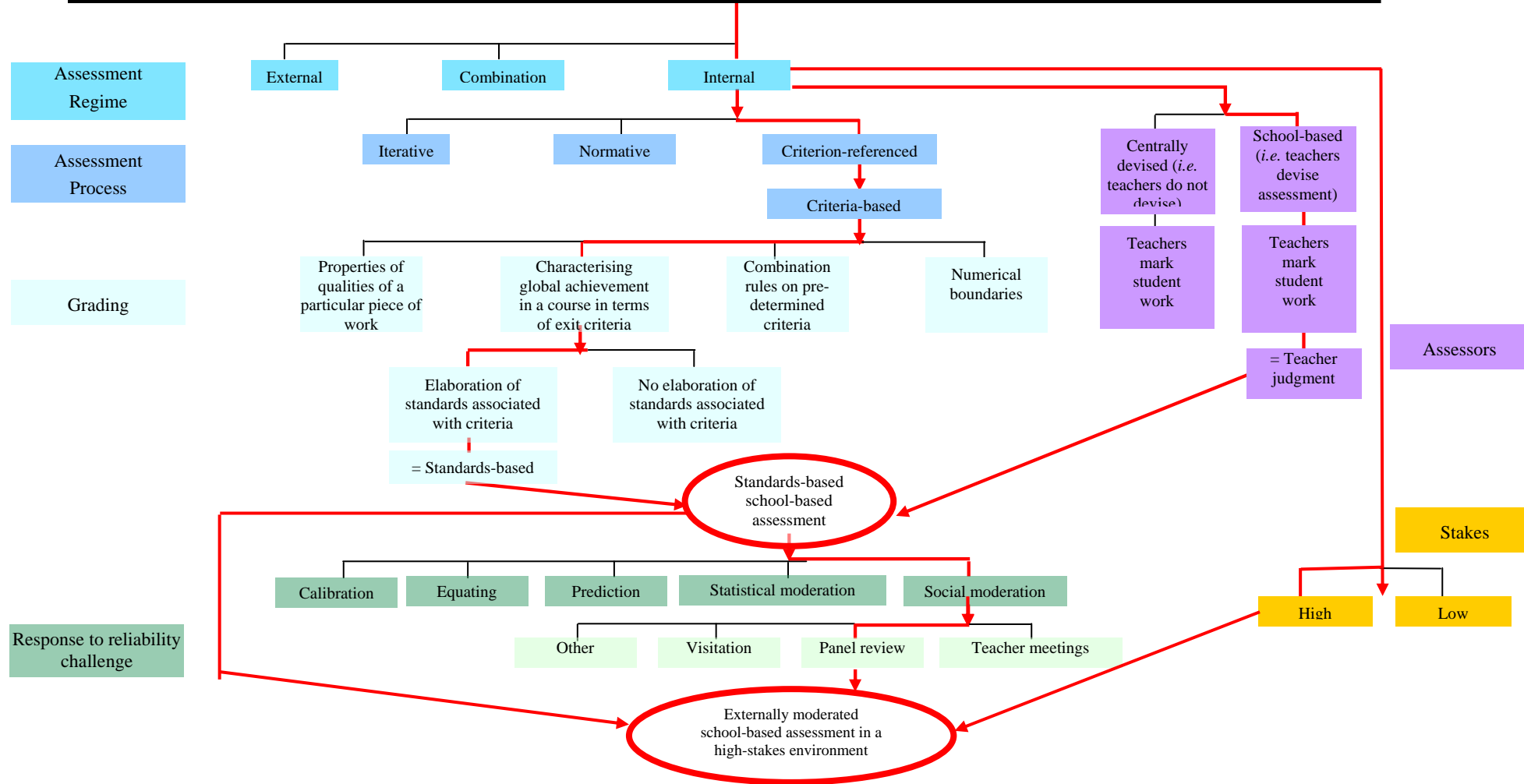


Figure 2: Tracing the route that led to the current Queensland senior system

References

- Beasley, W. (1986). *A Pathway of Teacher Judgments – From Syllabus to Level of Achievement* (Discussion paper 15). Brisbane: Assessment Unit, Board of Secondary School Studies.
- Beasley, W. (1986). *Profiling Student Achievement* (Discussion paper 17). Brisbane: Assessment Unit, Board of Secondary Schools Studies.
- Campbell, W.J., Archer, J., Beasley, W., Butler, J., Cowie, R., Cox, B., Galbraith, P., Grice, D., Joy, B., McMeniman, M., Sadler, R., & Wilkes, R. (1983). *Implementation of ROSBA in schools*. Unpublished report to the Minister for Education, Brisbane.
- Campbell, W. J., Bassett, G. W., Campbell, E. M., Cotterell, J. L., Evans, G. T., & Grassie, M. C. (1975). *Some consequences of the Radford Scheme for schools, teachers, and students in Queensland*. Final Report of Project. Brisbane: Australian Advisory Committee for Research and Development in Education.
- Clarke, E. (1987). *Assessment in Queensland Secondary Schools: Two decades of change 1964–1983*. Brisbane: Department of Education.
- Clarke, E. (1990). *Assessment in Queensland Secondary Schools: 1983–1990*. Brisbane: Department of Education.
- Elwood, J. (2006). Formative assessment: possibilities, boundaries and limitations. *Assessment in Education: Principles, policy and practice*, 13(2), 215–232.
- Fairbairn, K., McBryde, B., & Rigby, D. (1976). *Schools under Radford: A report on aspects of education in secondary schools in Queensland since the introduction in 1971 of school-based assessment*. Brisbane: Department of Education.
- Findlay, J. (1986a). *Improving the Quality of Student Performance through Assessment*. (Discussion paper 18). Brisbane: Assessment Unit, Board of Secondary Schools Studies.
- Findlay, J. (1986b). *Principles for Determining Exit Assessment*. (Discussion paper 21). Brisbane: Assessment Unit, Board of Secondary Schools Studies.
- Gipps, C. (1996). Assessment for learning. In A. Little & A. Wolf (Eds.), *Assessment in transition: Learning, monitoring and selection in international perspective*. Oxford: Pergamon.
- Gipps, C. & Stobart, G. (2003). Alternative assessment. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International Handbook of Evaluation* (pp. 549–576). Dordrecht: Kluwer Academic Publishers.
- Gunn, S. (2007). *Review of the literature on teacher judgment, standards, moderation and assessment*. Unpublished manuscript. Brisbane: Griffith University.
- Harlen, W. (2004a). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. *Research in Evidence in Education Library*, 3. Available at : <http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=54&Search=Harlen+2004>
- Harlen, W. (2004b). A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes. *Research in Evidence in Education Library*, 4. Available at: <http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=54&Search=Harlen+2004>
- Linn, R. L. (1993). Linking results of distinct assessment. *Applied Measurement in Education*, 6(1), 83–102.
- Masters, G.N., Forster, M., Matters, G.N., & Tognolini, J. (2006). *Australian Certificate of Education: A way forward*. Melbourne: Australian Council for Educational Research.

- Matters, G. N., Pitman, J. A., & O'Brien, J. E. (1998). Validity and reliability in educational assessment and testing: A matter of judgment. *Queensland Journal of Educational Research*, 14, 57–88.
- Matters, G. N. (2006a). *Assessment approaches in Queensland senior science syllabuses*. Report commissioned by the Queensland Studies Authority. Brisbane: Australian Council for Educational Research.
- Matters, G. N. (2006b). *Using data to support learning in schools: Students, teachers, systems*. Melbourne: Australian Council for Educational Research.
- Matters, G. N. (2006c). *Statistical moderation and social moderation around Australia*. Paper presented at the 32nd annual conference of the International Association for Educational Assessment (IAEA), Singapore.
- Matters, G.N., & Masters, G.N. (2007). *Year 12 Curriculum Content and Achievement Standards*. Brisbane: Australian Council for Educational Research.
- Maxwell, G. S. (2001). *Are core learning outcomes standards?* Brisbane: Queensland Studies Authority.
- Maxwell, G. S. (2007). *Implications for moderation of proposed changes to senior secondary school syllabuses*. Paper commissioned by the Queensland Studies Authority. Brisbane: Queensland Studies Authority.
- McMeniman, M. (1986a). *A standards schema* (Discussion paper 3). Brisbane: Assessment Unit, Board of Secondary Schools Studies.
- McMeniman, M. (1986b). *Formative and summative assessment: a complementary approach* (Discussion paper 6). Brisbane: Assessment Unit, Board of Secondary School Studies.
- McMeniman, M. (1986c). *Improving the quality of student performance through assessment* (Discussion Paper 15). Brisbane: Queensland Studies Authority. Available at: http://www.qsa.qld.edu.au/publications/yrs11_12/assessment/rosba015.pdf
- Myford, C.M. (1999). *Assessment for accountability vs assessment to improve teaching and learning: Are they two different animals?* Paper presented at ACACA conference Perth.
- Naughton, R. National curriculum journey begins. Office of the National Curriculum Board rose.naughton@ncb.org.au 2/5/08
- Pitman, J. A., & Herschell, P. (2002). *The Senior Certificate: A new deal*. Brisbane: Education Queensland and Board of Senior Secondary School Studies.
- Pitman, J.A. & O'Brien, J.E.(1999). *High-quality assessment: We are what we believe we do*. Paper presented at the IAEA Conference, Bled, Slovenia, May, 1999.
- Pitman, J. A., O'Brien, J. E., & McCollow, J. E. (2002). *A system combining school-based assessment and cross-curriculum testing considered in the context of some key debates in the educational assessment literature*. A paper in preparation of Pitman, J, A., 2002.
- Queensland Studies Authority. (2007). *Assessment: Years 11 and 12*. Available from: www.qsa.qld.edu.au/yrs11_12/assessment/index.html
- Queensland Studies Authority (2008). *Queensland Curriculum Assessment Reporting (QCAR) Framework*. Brisbane: Author.
- Radford, W.M.C. (1970). *Public examinations for Queensland secondary schools 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 students' achievements*. Brisbane: Department of Education.
- Sadler, D.R. (2000). *Assessing and grading: How we should rethink policy and practice*. Keynote address, Australian Universities Teaching Committee Forum, Canberra.

- Sadler, D.R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education*, 5(1), 77–85.
- Sadler, D.R. (1993). *Comparability in school-based assessment in Queensland secondary schools*. Brisbane, Queensland: Tertiary Entrance Procedures Authority.
- Scott, D.R., Berkeley, G.F., Howell, M.A., Schuntner, L.T., Walker, R.F., & Winkle, L. (1978). *A review of school-based assessment in Queensland secondary schools*. Brisbane: Board of Secondary School Studies.
- Shavelson, R.J., Black, P.J., Wiliam, D., & Coffey, J. (2004). *On linking formative and summative functions in the design of large-scale assessment systems*. Retrieved 3 May, 2007, from: http://www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/On%20Aligning%20Formative%20and%20Summative%20Functions_Submit.doc.
- Smith, C.M. (1995). *Teachers' Reading Practices in the Secondary School Writing Classroom: A reappraisal of the nature and function of pre-specified criteria*. Unpublished PhD Thesis. Brisbane: The University of Queensland.
- Strachan, J. (2002). Assessment in change: Some reflections on the local and international background to the National Certificate of Educational Achievement (NCEA). *New Zealand Annual Review of Education*, 11, 245–262.
- Tognolini, J. (2005). *Measurement*. Unpublished manuscript. Sydney: Australian Council for Educational Research.
- University of Queensland. (1978). Minutes of a special meeting of the professorial board. *The effects of the Radford Scheme*, G. Davies, 29 May 1978.
- Viviani, N. (1990). *The Review of Tertiary Entrance in Queensland 1990*. Report submitted to the Minister for Education. Brisbane: Department of Education.
- Wiggins, G. (1991). Standards, not standardisation: Evoking quality student work. *Educational Leadership*, 48(5), 18–25.
- Wyatt-Smith, C.M., & Cumming, J. J. (2003). Curriculum literacies: Expanding domains of assessment. *Assessment in Education*, 10(1), 47–60.
- Wyatt-Smith, C.M., Cumming, J., & Elkins, J. (2005). Opportunity to enhance learning: Redesigning assessment to sustain student engagement in the middle years. In D. Pendergast and N. Bahr (Eds.), *Middle Years Reform*. Allen & Unwin.
- Wyatt-Smith, C. M., & Matters, G. N. (2007). *Proposal for a new model of senior assessment*. Paper commissioned by the Queensland Studies Authority. Brisbane: Australian Council for Educational Research and Griffith University.

Appendix 1: Behind the abolition of public examinations

Comparison of 'pass' rates, selected subjects, 1962 and 1966

Subject	% of candidates receiving A, B or C	
Senior	1962	1966
English	88	85
Latin	84	75
Geography	79	68
Economics	71	58
Maths I	71	62
Chemistry	80	60
Physics	71	56

Distribution for all subjects, Senior, 1967

Grade	Description	% of candidates
7	Excellent	2-6
6	Very good	4-12
5	Good	10-20
4	Very fair	30-50
3	Fair	10-20
2	Unsatisfactory	4-12
1	Very unsatisfactory	0-6

Physics, Senior, 1967

Grade		% of candidates
7-4	'Pass'	30

Appendix 2: How the developing system documented itself

Discussion papers (Assessment Unit, Board of Secondary School Studies, 1986–1987)

1. *ROSBA's Family Connections*
Royce Sadler
2. *The Case for Explicitly Stated Standards*
Royce Sadler
3. *A Standards Schema*
Marilyn McMeniman
4. *Defining Achievement Levels*
Royce Sadler
5. *Subjectivity Objectivity and Teachers' Qualitative Judgments*
Royce Sadler
6. *Formative and Summative Assessment: A Complementary Approach*
Marilyn McMeniman
7. *Mathematics Criteria for Awarding Exit Levels of Achievement*
Janice Findlay
8. *Developing an Assessment Policy within a School*
Royce Sadler
9. *General Principles for Organising Criteria*
Royce Sadler
10. *Affective Objectives Under ROSBA*
Royce Sadler
11. *School-based Assessment and School Autonomy*
Royce Sadler
12. *Towards a Working Model for Criteria and Standards under ROSBA*
Marilyn McMeniman
13. *Criteria and Standards in Senior Health and Physical Education*
Robert Bingham
14. *Improving the Quality of Student Performance through Assessment*
Janice Findlay
15. *A Pathway of Teacher Judgments – From Syllabus to Level of Achievement*
Warren Beasley
16. *Assessment of Laboratory Performance in Science Classrooms*
Warren Beasley (with Peter Stannard, Burnside State School)
17. *Profiling Student Achievement*
Warren Beasley
18. *Principles for Determining Exit Assessment*
Janice Findlay
19. *Issues in Reporting Assessment*
Janice Findlay
20. *The Place of Numerical Marks in Criteria-Based Assessment*
Royce Sadler

IAEA Conference papers, other conference papers and occasional papers

Criteria-based assessment: The Queensland experience

JA Pitman, RP Dudley, 1985

Continuous quality control in written expression

JR Allen, 1988

The secondary-tertiary interface: the need for a new perspective

JA Pitman, GN Matters, 1989

The validity–reliability trade-off

GN Matters, JA Pitman, 1993

The Queensland Core Skills Test: Implications for the Mathematical Sciences
GN Matters, KR Gray, 1993

The Queensland Core Skills Test: In profile and in profiles
JA Pitman, GN Matters, 1994

An exploration of validity and reliability that tolerates distinct privileges: Standardisation and contextualised judgments
GN Matters, JA Pitman, JE O'Brien, 1995

Internal-consistency reliability measures for a test comprising three modes of assessment: Multiple choice, constructed response, and extending writing
JE O'Brien, JA Pitman, 1996

Using a hermeneutic technique to assess validity
JR Allen, EJ Bell, 1996 (AERA)

Are Australian boys underachieving?
GN Matters, JA Pitman, KR Gray, 1997

A case for testing generic skills
JA Pitman, GN Matters, NA Nuyen, 1998

Civitas: Participative research in the formation of Queensland school-based assessment 1970–1998
EJ Bell, 1998

Framing the future
EJ Bell, RP Dudley, 1999

New School Form, Forms and Benches: The move to benchmarks and outcomes
JR Allen, 1998

Meat Pies and Noodles: A vision of the future Australia-Asia-Pacific educational nexus
NA Nuyen, 1998

High-quality assessment: We are what we believe and do
JA Pitman, JE O'Brien, JE McCollow, 1999

Assumptions and Origins of Competency-Based Assessment: New challenges for teachers
JA Pitman EJ Bell, IK Fyfe, 1999

What young people say: Messages for educators
JA Pitman, 2001

Rich Task assessment in New Basics
GN Matters, 2004

The Queensland Assessment Task: Inventive authentic assessment designed to engage early adolescents of all ability levels
JA Pitman, KM Harris, 2005

Statistical and social moderation around Australia
GN Matters, 2006

Research papers, Board of Senior Secondary School Studies

- Some In-Practice Issues in Deciding SAIs: Case studies of six schools, 1999
- A Survey of Assessment Management Practices, 1992
- Comparability of Assessment in Mathematics Research Project Phase I, 1997
- Issues in Alternative Assessment in Mathematics A, B and C, 1997
- Studying Assessment Practices: A resource for teachers in schools, 1995
- Elements of Effective Curriculum Evaluation: An analysis of a Queensland model of curriculum evaluation, 2000
- Ideologies, Principles and Philosophies Underpinning Moderation in Australia, 1991
- Moderation of Achievements in School-Based Assessment, 1998
- Study of the Variability of Review Judgments: Stage 1, 1991
- Study of the Variability in Judgments of Review Panellists: Stage 2 – Accounting, 1995
- The Form R6: Content, language and length of comments and the level of school conformity with panel advice, 1994

- Popular Construction of the SEP, 1994
- Changing Populations and Changing Results: Gender differences in senior studies (1985–95), 1996
- Language and Equity: A discussion paper for writers of school-based assessment instruments, EJ Bell, N Simpson, 1995.
- Report on the Survey of Schools as Part of the Development of Board-Registered Syllabuses, 1996
- Random Sampling 1999 – a quality assurance process for moderated school-based assessment, 1999
- The development and approval of syllabuses for Board subjects, RP Dudley, DA Abbey, 1999
- Making Subject Choices – Research into factors that influence students in the selection of senior secondary subjects, 1998
- Assessment of higher order thinking skills: A discussion of the data from the 2001 random sampling exercise and a workshop for teachers, EJ Bell, JR Allen, P Brennan
- Issues and Practices in Deciding Competency, DJ Kelly, EJ Bell, 2000
- Elements of Effective Curriculum Evaluation
- Student Participation and Student Outcomes in the Social Sciences, J Williams, EJ Bell

Reviews, journal articles

(selection involves one author of this paper in a previous role; there are many other papers/authors)

Tertiary Entrance in Queensland: A Review, 1987 (JA Pitman, Chair of the Ministerial Working Party)

A design process for constructing the QCS Test, GN Matters, 1991

The Queensland Core Skills Test: Evaluation of Design Criteria and Process, G Trost, 1992

The Queensland Core Skills Test: A follow-up evaluation of design criteria and process, G Trost (Bonn) 1996

Meeting the challenges of short-response items on the QCS Test, GN Matters, 1991

A report of the scan of the Queensland senior curriculum to identify the common elements, JR Allen, GN Matters, RP Dudley, & PK Gordon, 1992

A report on the feasibility and utility of item banks, GN Matters, 1994

Gray, KR, & Matters, GN (1994). Principles of combinatorial design exploited in the 1993 QCS marking operation. *Utilitas Mathematica*, 48, 33–64.

Matters, GN, Allen, JR, Gray, KR, & Pitman, JA (1999). Can we tell the difference and does it matter? Differences in achievement between girls and boys in Australian senior secondary education. *The Curriculum Journal*, 10(2), 283–302.

Matters, GN, & Burnett, PC (1999). Multiple-choice versus short-response items: Differences in omit behaviour. *Australian Journal of Education*, 43(2), 117–28.

Matters, GN, & Burnett, PC (2003). Psychological predictors of the propensity to omit short-response items on a high-stakes achievement test. *Educational and Psychological Measurement*, 63(2), 239–56.

Matters, GN, & Gray, KR (1995). The Queensland Core Skills Test: Implications for the mathematical sciences. *Unicorn*, 21(4), 74–89.

Matters, GN, Pitman, JA, & O'Brien, JE (1998). Validity and reliability in educational assessment and testing: A matter of judgment. *Queensland Journal of Educational Research*, 14(2), 57–78.

Pitman, JA, & Matters, GN (2000). An approach to cross-curriculum testing (NCME International News). *Educational Measurement: Issues and Practice*, 19(4), 25–26.

NB: The above list is not exhaustive. It does not include handbooks, primers, guidelines and so on. Nor does it sample adequately across the various sections of the Board's operations.

