A report to the Queensland Studies Authority

Assessment approaches in Queensland senior science syllabuses

Gabrielle Matters
Australian Council for Educational Research

August 2006

1. Introduction

This paper deals with appropriate assessment in senior science subjects in Queensland. It was commissioned in response to an apparent jumble of attitudes surrounding the application of criteria-based assessment to the new science syllabuses.

1.1 Terms of reference

- To investigate the theoretical underpinnings of the two approaches to criteria-based and standards-referenced assessment in the Queensland senior science syllabuses;
- To make recommendations as to which approach is the more applicable in the current context;
- To model the proposed approach, and provide advice about implementation.

See Appendix 1 for a copy of the letter of appointment.

1.2 Sources of information

Initial briefings were received from Mr Peter Luxton (Deputy Director, Curriculum, QSA). Hard copies of the QSA senior science syllabuses were acquired and studied (see list below) and the QSA website was visited. Permission for access to work programs, one in each of Chemistry, Physics and Science21 was granted; these three work programs came from three different trial-pilot schools. It was also made possible to monitor traffic on the closed discussion list set up for trial-pilot schools. QSA's policy statement on exit assessment was re-visited. The educational measurement literature was scanned. Research findings and personal experience gleaned within the Australian education sector and in working with education systems overseas was also drawn upon. Some descriptions of the Queensland system of externally moderated school-based assessment ('the Queensland senior system'), which had, of necessity, been composed by the author of this report to counter 'bad press' elsewhere were included as appropriate.

Deliberations of Science Subject Advisory Committee meetings, QSA's Curriculum Committee meetings and the QSA Governing Body meetings were not taken into account. This paper deals only with the outputs of QSA, which, to a certain extent, are inputs to schools and teachers and students for curriculum and assessment as they work in partnership with QSA.

1.3 Queensland senior science syllabuses

The suite of science subjects offered to students in Years 11 and 12 in Queensland comprises Chemistry, Physics, Biology, Earth Science and Multi-Strand Science. Science21 is a new science subject being trialled. (It is acknowledged that Agricultural Science and Marine Studies are subjects with a scientific basis. They are not discussed in this paper.)

All of these subjects have the status of Authority subject; that is, they go through the full gamut of QSA's moderation procedures and can 'count' in the calculation of a student's tertiary entrance rank – which, in Queensland, is the Overall Position (OP).

Currently, there is only one incompatibility rule that relates to the senior science subjects: 'A student's Senior Certificate will record achievement in Multi-Strand Science plus one only of the following senior science subjects – Chemistry, Physics, Biological Science, Earth Science – in any one semester'. Multi-Strand Science, as its name implies, contains elements of all the other sciences.

The current syllabus documents carry the following (unedited) imprints:

> *Chemistry senior syllabus 1995*
> This syllabus is approved for general implementation until 2001 unless otherwise advised.
>
> *Chemistry Extended Trial Pilot Syllabus September 2004*
> Restricted Use Approved Schools Only
>
> *Physics senior syllabus 1995*

This syllabus is approved for general implementation until 2001 unless otherwise advised.

Physics Extended Trial-pilot Syllabus September 2004
Restricted Use Approved Schools Only

Biology Senior Syllabus 2004

Multi-Strand science: senior syllabus 1998
This syllabus is approved for general implementation until 2005 unless otherwise stated.

Earth Science 2000
This syllabus is approved for general implementation until 2006 unless otherwise specified.

There is also a document labelled thus:
Science21 (working title) Trial Senior Syllabus 2004
To be used in approved schools commencing with Year 11 Students only in 2005
Science State Smart State Initiative.

The investigation reported in this paper does not include Earth Science (234 students in Years 11 and 12 in 2005).

1.4 Summary of the existing situation

For each of Chemistry and Physics, there is an 'old' syllabus (from 1995) that remains the current syllabus (in 2006, beyond the 2001 date mentioned above) and a 'new' syllabus (from 2004) that is currently being evaluated as a result of its implementation in trial-pilot schools (121 for Chemistry and 104 for Physics) over 2005–06. For both Chemistry and Physics, the major concern is that the new syllabuses demand a significantly different approach to assessment. This approach is portrayed as holistic grading versus analytic grading. The issue of holistic grading is explored in this paper.

For Biology, the 'new' syllabus (from 2004) is being generally implemented; that is, an evaluation was considered to be unnecessary. Science21, a product of the Science State Smart State Initiative, is currently being evaluated as a result of its implementation in 45 trial-pilot schools. For Biology, the concern is the alternative on offer through Science21 (and also Multi-Strand Science, although Science21 is a potential replacement for Multi-Strand Science). The task-based approach of Science 21 could be portrayed as the pursuit of authentic assessment. Authentic assessment and task-based curriculum/assessment are explored in this paper.

In most of what follows, it is the new senior science syllabuses, Chemistry, Physics and Science21, whose assessment approaches are under investigation. Some of the criticisms (not that this report is a litany of critiques) can be taken to apply to the old syllabuses as well as the new syllabuses, and to other subjects as well as the sciences. Arguably, it is an appropriate time to ask the general question about the current state of play in assessment of senior subjects in Queensland.

1.5 Major issues

The current challenges for science teachers posed by the new syllabuses have been identified (QSA, 2006) as follows:

- Meeting the requirement for a variety of assessment strategies;
- Using task-specific descriptive standards in the place of marks;
- Generating integrated assessment tasks (for holistic grading);
- Making holistic judgments about standards;
- Awarding exit levels of achievement for certification in line with stated policy, and generating finer-grained data for calculation of tertiary entrance ranks.

See Appendix 2 for a copy of the letter providing background and scope.

The language used to articulate the current challenges is itself challenging, and underscores fundamental problems that now declare themselves. They will be addressed in the body of this report.

It is acknowledged at the outset that these issues are confronting. Designing valid assessment instruments that produce reliable results is not simple and involves many more hours than a ‘rainy Sunday afternoon’ to fulfil the promise of internal assessment. Finding workable ways to set, disseminate and maintain standards is not trivial and involves many people in the ‘guild of professionals’ in extended dialogue and experimentation.

A huge amount of intellectual capital has already been expended to date – in translating the intentions of the old science syllabuses into student experiences and in developing new science syllabuses with an eye to the futures agenda (the future that is already upon us), with an eye to the relevance push (especially in Science where the number of students participating is not proportional to the needs of our nation or the world), and with every good intention to put into practice the very best models of assessment.

It should be noted at the outset that the QSA senior syllabuses possess all the components of a rigorous curriculum framework as defined in the education literature, they are comprehensive in their coverage (from content prescription to equity statement), and they have been studied and deemed worthy of note by educationists in other countries. It is the assessment challenge that definitely requires continuing attention.

2. Other big issues and an early conclusion

Associated with the major issues identified by QSA (and listed in Section 1.5) are other (big) issues that also require attention. These other issues are not easy to treat in isolation from each other because, to a certain extent, each one shapes the other. The four big issues are:

1. The fact that the Queensland system of criteria-based assessment developed, not so much underpinned by theory but more so as a theory-building exercise in itself;
2. The derivation over time of myriad interpretations of the traditional criteria/standards matrix;
3. The apparent confusion in notions of assessment regime, assessment criterion, assessment technique, and assessment instrument[1];
4. An over-emphasis on the distinction between holistic and analytic, together with the absence of a convincing reason for introducing a dichotomy in the first place.

The question also needs to be asked as to whether there has been a fair portrayal of science teachers in the education community, given that they are often caricatured as troglodytes attached to marks and numbers. The long tradition of education has all teachers ‘marking’ (as witnessed in their conversations about marking loads). A confident approach to educational assessment is gained through familiarity with numbers, scales and measurement. This is not to deny that being fixated on (perceived) objectivity could signal an absence of workable alternatives but it does highlight an essential point: Teachers have to make judgments about qualities and quality. Sometimes counting or measurement will form an essential step in determining quality or competence. On the other hand, too great a preoccupation with numbers and scores may get in the way of determinations of quality (Sadler, 1986).

In this paper, each of the four issues listed above is discussed in turn despite the high degree of interrelatedness between them. Section 3 looks at theoretical underpinnings, Section 4 at the traditional criteria/standards matrix, and Section 5 at notions of assessment including analytic and holistic rubrics and standards-referenced assessment.

A reader of this report who is familiar with the topics in Sections 3, 4 and 5 might elect to skim those sections.

[1] Even assessment *strategy*, as mentioned in major issues (Section 1.5)

Section 6 applies the theory and critique from earlier sections to criteria and standards in the new syllabuses, Section 7 presents a check list for setting, describing and representing standards, Section 8 reviews important information about other ways of assigning grades, Section 9 deconstructs the existing model for presenting criteria and standards, and Section 10 provides a reconstruction in the form of a new standards schema.

This examination leads to the conclusion that, despite all the good intentions that no doubt accompanied the development of new syllabuses and new approaches to assessment, the situation has become too complicated and too far removed from a few simple concepts. The argument that threads this paper is for the essence of standards-based assessment. Also, a case is put for not mandating the incorporation of task-based assessment into new science courses even though authentic assessment is *de rigueur* in modern assessment practice. This does not imply that task-based assessment should not continue to be explored and employed.

Conclusions and recommendations appear in Sections 11 and 12.

Implicit in all of the discussion are notions of validity and reliability, the cornerstones of educational measurement.

3. Theoretical underpinnings

3.1 Where are the references and referents?

In moving from a system of external examinations to school-based assessment in 1972 (see Radford, 1970) and then from norm referencing to criteria-based assessment in 1980 after the implementation of the Scott Report (Board of Secondary School Studies, 1978), Queensland was seen to be quite radical. As there were few examples, if any, of criteria-based school-based assessment for certification purposes, Queensland was unable to be derivative of other systems even if it had wanted to be so. A version of criterion-referenced assessment ‘morphed’ into, and became known as, ROSBA[2], the very title indicative of its genesis. To this day, Queensland and the ACT are the only jurisdictions (out of eight) in Australia where no external examinations are administered to Year 12 students and where, therefore, teacher judgments, neither combined with, nor scaled[3] to, external subject-specific examination marks, are reported on the Senior Certificate and recognised by universities and employers.

Other jurisdictions in Australia are gradually coming to realise the value of internal assessment[4], as are other countries (e.g. the UK and South Africa).

It is thus understandable that the Queensland senior assessment system is not so much underpinned by theory as having been and continuing to be a theory-building exercise in itself. The result of a preliminary Google search of the World Wide Web highlights this point: the search for ‘criteria-based assessment’ yielded 669,000 references, the top of the list being ‘QSA: Years 11 and 12: Assessment: Criteria-based assessment’. A further search for ‘criteria-based assessment Queensland’ yielded 29,700 references, predominantly a subset of the original 669,000 (the same QSA reference being at the top of the list). The third entry was for the Pandora Archive – and it yielded Pitman and Dudley’s paper on the Queensland experience of criteria-based assessment that was delivered at the 1985 IAEA (International Association for Educational Assessment) conference in Oxford (this was not long after IAEA had been established by luminaries in educational measurement from the Educational Testing Service (ETS), Princeton, and conference papers were awash with reports on standardised testing). In 1987, Sadler’s seminal paper on ‘specifying and promulgating achievement standards’ was published in the *Oxford Review of Education*.

[2] Review of School-Based Assessment

[3] Some writings give the impression that the QCS Test is used to validate teacher judgments (as in statistical moderation). It is not. Social moderation is used to validate teacher judgments. The QCS Test is used to put results from different subject-groups and different schools onto the same scale for tertiary entrance ranking of students.

[4] Externally moderated school-based assessment

Papers written by Sadler, McMeniman and Beasley in the 1980s are the stuff of theory building: Much of the terminology that is used now was introduced in that era, and many of their papers are still cited today. The senior system in Queensland thus became and remains a referent for other parts of the world. Here are some cosmopolitan fragments:

From the UK, from the editorial (1999) of a refereed journal, *Assessment in Education: Principles, policy & practice*:

> Queensland: one of the parts of the world where the most sustained attention has been given to providing high-quality school-based assessment.

From the US, from a keynote address (2000) given by Dr Carol Myford, from ETS:

> 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 it 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.

These examples go some way to explaining why the reference list for this paper is peppered with the names of Queenslanders. The disproportionate amount of referencing to Queenslanders is neither nostalgia nor chauvinism on the part of the author of this report.

One factor contributing to the success of the Queensland system is that it sends powerful messages to teachers, students, parents and the wider community about what really counts through five key elements. These five key elements of an effective assessment system underlie the best of a wide range of assessment policies and practices that permeate the educational research literature (J. R. Allen, 2003, personal communication). They are stated below in general terms with Queensland-specific terms following in square brackets:

1. There are guidelines that teachers/schools must use in planning [syllabus].
2. There are formal plans for student learning and achievement that teachers/schools must make [work program].
3. Evidence of student achievement must be produced [folio of student work].
4. This evidence must be assessed against the guidelines and plans [teacher judgment based on pre-set standards].
5. There is a process for validating teacher judgments of student achievement [social moderation].

The system scores 5 out of 5, the hallmark of a coherent and effective system (Matters, 2004) … and this is far from universally the case. Particular components of this assessment system are in line with requirements for obtaining consistent and valid moderation; namely, standards, evidence and consensus (Pitman, O'Brien, & McCollow, 1999). And, in pointing very directly to questions of curriculum intent, the 5-element approach has particular benefits for task-based curriculum/assessment as proposed in QSA's new science syllabuses.

3.2 A rose by any other name?

The next discussion is not mere semantics. The intricacies are essential for a proper appreciation of the theoretical elements of criteria-based assessment in Queensland and of the theory building that has occurred over the past 25 years of practice and that has been expanded upon, and complemented by, theoretical positions developed outside Australia.

Some assertions as a starting point

The problems associated with the ambiguity of the term *criteria* explain some of the current jumbled attitudes to criteria-based assessment in the senior sciences (and possibly in other subjects).

The philosophy, policies and procedures in the senior system refer to a system that is actually *standards-based* (compared with standards-referenced or criteria-based).

If one accepts the proposition that the senior system is both criteria- and standards-based, then holistic grading and standards-based assessment are *not complementary*. If one accepts the proposition that the senior system characterises a student's global achievement in a 2-year course of study in terms of exit criteria only (i.e. no elaboration of standards), then holistic grading and criteria-based assessment *might be complementary* (but is a system with no detailed elaboration of standards really desirable?)

Analytic marking is primarily a technique for dealing with trade-offs across assessment criteria.

Some alternative approach to the criteria/standards matrix might be capable of accommodating all the needs of the apparently different approaches.

The meaning and use of the terms *standards-based*, *holistic and analytical* are discussed within Sections 4 and 5.

4. Criteria- and standards-based assessment

Sadler (2000) provides two reasons for the problematic nature of the use of the term *criterion*:

1. The term *criterion* is often used when *standard* is meant.
2. The term *criteria* has multiple meanings.

The problem is now analysed.

4.1 Criteria and standards distinguished

In everyday conversations, the words criteria and standards are often used interchangeably even though they are defined differently in the dictionary. Here are the dictionary meanings of these two key terms.

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.

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.

Here are the meanings ascribed to these two key terms in the educational measurement literature.

Criteria (criterion, *sing*.): Those properties, dimensions or characteristics by which student performance[5] is appraised; for example, sensible motivation, data presentation, and scientific analysis are three criteria that could be used in assessing project work on a student's personal health plan.

Standard (assessment standard): Fixed point along the criterion describing/representing qualitative (and discernible) differences in student performance. Standards are the referents that underlie judgments about success or level of merit in a performance (see Maxwell, 2001).

For the student, the highest standard is a goal to aim for. For the teacher, standards are used in assessing or describing the quality of student performance. S/he judges which one of several designated standards best represents the characteristics of a student's performance; that is, what label

[5] Throughout this paper, the terms *student performance* and *student achievement* are used interchangeably

to attach to the performance or what category (such as A–E) in which to place it. (The term *category* is juxtaposed with *continuum* in Section 9.1.)

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 (see Section 7).

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.

The writings on the closed discussion list are commendable in that notion of weighting is deemed inappropriate in the present discourse.

Combining judgments on separate criteria involves trading off.

4.2 Trade-offs and 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 (as described in Section 4.1). Either way, the notion of trading off or making on-balance judgments is crucial to deciding a student’s exit level of achievement in the senior sciences in Queensland.

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? Section 5.9 describes five models for dealing with trade-offs.

4.3 Criteria-based assessment and standards-referenced assessment compared

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-referenced assessment or both?

Queensland’s senior assessment system can be 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 presuppose criteria.

This leads to an appreciation of why Queensland’s senior assessment system can be called standards-based. In their purest form, standards are descriptions or other specifications of performance levels that are free from any references to the performance of the typical student, the proportion of students expected to achieve a given level, or the particular age or stage of schooling at which a certain level of performance is thought to be acceptable.

But can the system be called standards-referenced? 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 NSW 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.

For the remainder of this paper, the Queensland system is taken to be both criteria-based and standards-based, a local variety of criterion-referenced assessment, the architects of the system having drawn on the limited literature and experience elsewhere, and then created their own operating system. In particular, the Queensland system is concerned with specifying what the various levels of achievement are to consist of, taking into account the various types of objectives as stated in the corresponding syllabus document.

In Queensland, the terms *grade specifications* and *standards descriptors* seem to be used interchangeably. The difference, however, is that grade specifications specify (of course) what the various levels of achievement (grades) are to consist of whereas standards descriptors describe (of course) what student work for the award of particular levels of achievement (standards) is to be like.

A standards descriptor is 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. This could operate at the domain level or the task level or the examination level or on exit from a course of study.

To what species do the words in the cells in the criteria/standards matrices for the science subjects belong? What happens if the definitions above for standards descriptor and grade specification are applied to the words in the cells in the matrices in the new syllabuses? The words in the cells fit the definition of grade specifications even though they are actually called standards descriptors. This conclusion should be subject to verification and, if found to be true, studied as a possible explanation of why the required loop between general objectives and standards descriptors remains unclosed to a certain extent (see further discussion in Sections 7 and 11).

The standards descriptors (now looking more like grade specifications) reiterate the general objectives rather than complement them through a new set of words that refers to student work/performance. The features of student work required for the award of a given level of achievement are not fully described. The so-called standards descriptors for assigning an exit level of achievement are too general.

5. Terms and concepts

5.1 Assessment level, regime, criterion, technique, instrument

We can talk about assessment at the level of an individual task or examination question. We can talk about assessment at the level of a collection of tasks or questions on an examination paper. And we can talk about assessment on exit from a course of study. Many of the principles and arguments can apply at all three levels. Again we need to have a shared understanding of theoretical concepts that underpin assessment practice.

As mentioned in Section 2, this investigation revealed a lack of clarity surrounding the terms assessment technique, assessment instrument, assessment criterion, and assessment regime. The conflation of instrument and criterion might be another reason for the jumble of attitudes that have been identified in approaches to assessment in the senior science syllabuses.

Assessment *regime* refers to the established style (often under government control) for identifying, gathering and interpreting information about student learning. It might be via a standardised testing program or a suite of external examinations or, as in Queensland, internal assessment.

An assessment *criterion* is one of several levers by which judgments about the quality of student work are made and defended, student work being the outward and visible sign of the learning process. For example, the exit assessment criteria in the Science21 syllabus are: Knowledge and conceptual understanding, Working and thinking scientifically, and Impacts of Science.

Assessment *technique* refers to the method used to gather evidence about student achievement. Techniques include observation, consultation, and analysis of student work. The quality of student work, such as a piece of extended writing, as in a single essay, can be judged using analytic or holistic techniques. The new senior science syllabuses require holistic techniques for judging the worth of a 2-year collection (folio) of student work.

An assessment *instrumen*t is a tool or device or constructed situation that has the primary aim of bringing forth evidence about student achievement. It could be a written assignment, oral presentation, performance, demonstration of mastery, practical work, field work, test, examination, project, viva voce, rich task[6] and so on. The Science21 syllabus gives primacy to written tasks, extended investigation reports, collections of work, and non-written presentations.

The discourse surrounding the senior science syllabuses tends to demonise traditional assessment instruments such as formal tests and examinations and glorify large integrated tasks. This issue is revisited in Section 10.5 (task-based assessment/curriculum).

School-based assessment requires good work programs and good assessment techniques because criteria/standards-based assessment does not adjust for variability in assessment techniques and instruments. It is concerned with the criteria that can be used for judging the quality of student work and with explicit statements about standards. For this reason, students need to be ‘let into the secret’ about what standards they should aspire to. This is made difficult by statements of standards in syllabuses that are not clear and specific.

5.2 Grading and marking distinguished

The term *marking* is reserved for the process in which a teacher–assessor assigns a code to a piece of student work such as a response to an individual task or an end-of-semester test (e.g. B, 60/100). The term *grading* is reserved for the process in which a teacher–assessor (or group of teacher–assessors) assigns an exit level of achievement to a collection/folio of student work, say a portfolio of results obtained over a course of study. Grade is often used to mean the single result for reporting (i.e. after combining/aggregating results of several assessments).

Marks and grades are codes for the standard of student performance on a single assessment task/question/instrument and on a collection (course) of assessment tasks/instruments, respectively. There are various symbolic codes used, typically consecutive letters, with *A* denoting the grade pertaining to the highest standard of performance and the number of grades varying from task to task.

Grading student work can be difficult because judgment calls must be made. In theory, five different teacher–assessors could give five different grades to an individual student’s folio. Nevertheless, a study of the level of agreement between raters of student folios in Queensland’s moderation program concluded that there was ‘an exceptionally high level of agreement…between assessments. These levels of agreement are significantly higher than the levels…typically reported for independent assessments of student work – including independent assessments of external examinations’ (Masters & McBryde, 1994).

School-based assessment as practised in Queensland is much more than having teachers’ judgments count in a high-stakes environment – it has teachers setting the assessment instruments, composing the associated marking schemes, and deciding upon the conditions under which the assessment will occur. Given that social moderation is Queensland’s response to the reliability challenge that

[6] Used in the general sense to describe multidimensional, multi-modal activities completed over an extended period of time

inevitably accompanies school-based assessment, any changes to the assessment system require careful consideration in terms of potential consequences for reliability.

5.3 Marking rubrics as an assessment technique

Marking rubrics are descriptive marking schemes developed by teachers or other assessors ‘to guide the analysis of the products or processes of students’ efforts’ (Brookhart, 1999). Marking rubrics are typically employed when a judgment of quality of student work is required.

Theoretically, marking rubrics can be used for any year level and to assess a broad range of subjects and activities. They have been used in the assessment of extended writing, group activities, extended projects and oral presentations in the USA and elsewhere. In Queensland, marking rubrics have been used in subject-specific assessments at school level and in the assessment of QCS test items (short response and extended writing). Presumably, being upfront about the criteria and their relative importance would stop one marker from weighing heavily on linguistic structure while the other was more interested in the persuasiveness of the argument in the Writing Task (WT) response. Marking the WT proceeds on the basis that good writing would likely have a combination of these and other factors.

Marking rubrics are not mere check lists (for on/off decisions as in competency-based assessment or in yes/no non-discountable criteria). They are based on descriptive scales and support the judgment of the extent to which criteria have been met. In Australia, they might be called marking schemes, marking guides, criteria sheets and so on. The general term scoring rubric is used in this section.

There are different types of scoring rubrics – analytic and holistic, general and task-specific.

5.4 Analytic or holistic or impression

Brookhart (1999) provides the following definitions.

> *Analytic scoring rubrics* allow for the separate assessment of each of several (two or more) criteria. Each criterion is scored on a different descriptive scale.
>
> *Holistic scoring rubrics* support broader judgments concerning the quality of the process or product. In the holistic scoring rubric, the criteria are considered together on a single descriptive scale.

The criteria are not considered together on a single scale (or dimension or criterion) in the new senior science syllabuses even though the process is referred to as holistic grading.

Holistic/impression and analytic are traditionally about marking student scripts as opposed to grading folios of student work on exit from a course of study. The terminology of marking (rather than grading) is maintained in the discussion that follows.

The terms *impression marking* and *analytic marking* are most often used in the context of large-scale testing or marking operations where speed and reliability are paramount. It is claimed that analytic marking ensures intra-marker consistency and inter-marker reliability. It is claimed that impression marking can be very reliable if there is appropriate training before marking and continuous monitoring of markers during the marking operation.

Impression marking refers to situations where a marker, using his/her expert judgment, assigns a mark on the basis of an overall impression of the work’s worth. The word ‘impression’, with its overtones of impressionistic can give the wrong impression (no pun intended) about capturing the distinction in question. After all, an impression is, by definition, something that leaves a trace – such as the image of student work.

Impression marking is not unlike the classic wine-judging model. Separate (analytic) scores are not applied to three criteria (e.g. viscosity, nose, palate) and then aggregated. Rather, a well-informed judge trades these off (privately) and makes an overall judgment. ‘Chateau Cardboard’ will always get a lower score than Grange Hermitage, provided the wine judge is not a maverick.

If suitably interrogated, an impression marker should be able to give some account of why marks were assigned as they were. S/he will be following some sort of private marking scheme with associated weightings or internalised trade-off rules and priorities. A perceived problem with impression marking is that markers may differ considerably in their private marking schemes and so different markers could give very different grades to a given piece of student work.

Analytic marking is where some sort of marking scheme is employed that gives guidance to the marker about what features s/he should look for and what weighting should be given to them (or what influence each feature's mark should have on the overall mark).

Constructed-response items lend themselves readily to analytic marking as can be seen in the following example of an analytic marking scheme (original question not included).

Example of an analytic marking scheme

A total of 10 marks is allowed for this question.

5 marks for demonstrating understanding of the basic concepts of two theories

For theory 1, students should be rewarded for making specific mention of developmental stages, adaptation as assimilation and accommodation, schemata, sensory-motor operations, concrete operations, formal operations.

For theory 2, students should be rewarded for making specific mention of personal constructs and construing, validation and core constructs.

5 marks for noting significant differences between the two theories, in particular:

Theory 1 is a developmental theory.
Theory 2 grew out of a concern with therapeutic change.
Theory 1 accounts for motivation in terms of 'equilibration'.
Theory 2 accounts for motivation in terms of 'a search for meaning'.
Theory 1 is logico-mathematical; theory 2 is not.

A possible problem with analytic marking is that, if it is too specific, it may be difficult to differentiate between excellent students and competent students. In deciding the credit given (marks awarded) to different parts of the answer, it is important to allow some flexibility in the marks to allow discrimination; for example, by having bonus marks awarded for overall quality or originality.

Choosing an analytic scoring rubric does not eliminate the possibility of an holistic factor. An holistic judgment may be built into an analytic scoring rubric as one of the score categories. One difficulty with this approach is that overlap between the criteria set for the holistic judgment and the other properties being assessed cannot be avoided. When one of the purposes of the assessment is to assign a grade, this overlap should be carefully considered and controlled. The teacher–assessor (or syllabus writer) should determine whether the overlap results in certain criteria having more influence than was originally intended. In other words, the teacher–assessor needs to be careful that the student is not unintentionally penalised severely for underperformance on one dimension (or vice versa).

Sometimes it is impossible to separate a judgment into independent properties. When there is an overlap between the criteria identified for the assessment, an holistic scoring rubric may be preferable to an analytic scoring rubric.

Appendix 4 describes a research activity to investigate, *post hoc*, the dimension (or dimensions) that underpin holistic grading in the new science syllabuses.

5.5 General or task-specific or both

Scoring rubrics may be designed for the assessment of a specific Science task or the assessment of a broader category of tasks (e.g. all the tasks that make up a 2-year assessment program). If the purpose of a given course is to develop a student's knowledge and understanding of the science subject, a general scoring rubric may be developed and used to assess each of the tasks done by that student. This approach would allow the students to use the feedback they acquired from the last presentation to improve their performance on the next presentation – which complements the principle that fullest and latest information should be the basis for decisions about exit levels of assessment.

If each task focuses upon a different scientific concept or field of inquiry, and the purpose of the assessment is to obtain evidence of the students' knowledge of that particular topic, a general scoring rubric for assessing a sequence of tasks may not be adequate. Scientific topics or problems differ in both influencing factors and outcomes. In order to assess students' factual and conceptual knowledge of these topics it would be necessary to develop separate scoring rubrics for each task. This is called a task-specific scoring rubric because it is designed to judge the quality of student performance on a single assessment event.

Scoring rubrics can be designed to contain both general and task-specific components. If the purpose of a presentation is to assess students' inquiry skills and their knowledge of the scientific topic that is being investigated, an analytic rubric could be used that contains both a general component and a task-specific component. The investigative component of the rubric can consist of a general set of criteria developed for the assessment of inquiry skills; the task-specific component of the rubric can contain a set of criteria developed with the specific topic/phenomenon in mind.

The point being made here is that combining results on a series of tasks within a single course of study requires careful consideration of what is being required of students and what is being rewarded in the marking scheme (scoring rubric), and how the specific and general components of the results will map onto the features of student work that are described in the exit level statements.

5.6 Grading according to standards

By their very nature student folios allow for a variety of ways of having student work match the standard for an exit level of achievement; indeed each student folio is in some sense unique. Having pre-set standards descriptors and pre-determined trade-off rules reduces this variability by providing a limited set of meaningful codes to each student.

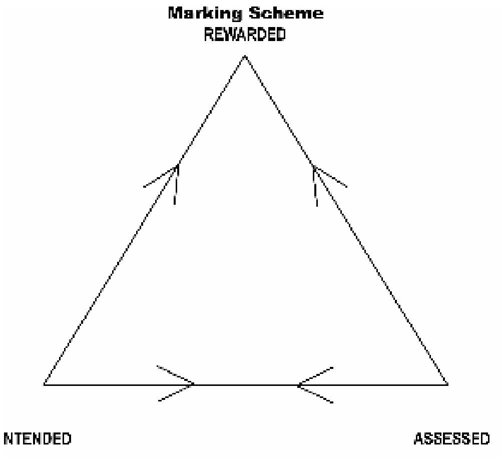
There are obvious problems of accountability and validity in the notion of teachers assigning a grade based on an impressionistic approach about the worth of a particular student folio. A tight mapping of the outcomes/objectives/topics being assessed in each task against the desirable features of student work at the highest exit level of achievement might improve this situation. It would provide a mechanism for ensuring that there is an appropriate balance across the collection of tasks. Students have a legitimate expectation that the choice of teacher (and therefore of tasks in a collection) should not affect the grade they are given for a subject.

To meet these expectations it is necessary to give teachers/schools clear and specific instructions in the syllabuses. Syllabuses should be sufficiently prescriptive to guide teachers/schools and yet not so detailed that they are useless in practice. It would be hard to argue that the new science syllabuses are sufficiently prescriptive about the issues being raised here. The syllabuses do contain large amounts of useful material for teachers/schools but, in my opinion, it is not of the type that delivers the discrimination between students that is vital in assigning Subject Achievement Indicators (SAIs) as input into the calculation of the OP.

The appropriateness of the scoring rubric as an assessment technique depends on the purpose of the assessment. The distinction is essentially between publicly and privately agreed ways of trading off. In my opinion, it is not necessary to mandate holistic grading even it is considered to be the fairest or most expedient way so long as the technique for grading is true to principles of standards-based assessment and trading off within that assessment model.

5.7 Three essential elements

In a good assessment task, there is an effective interplay of three simple elements – what is taught/learnt (the intentions of the curriculum), what is assessed (evidence of knowledge, skills/dispositions in the domain being sampled), and what is rewarded (high-quality performance on the criteria set down in the marking scheme and incorporated in an associated exemplar).



Syllabus Objective Assessment Task

Applications of element interplay

By studying the extract below, it can be seen that *what is assessed* by Question 27 on the 2004 HSC Physics Examination reflects the syllabus outcome statements (and we assume that *what is taught* in classrooms is an enactment of the intended curriculum). Also, the criteria in the marking guidelines reflect the commands in the examination question (e.g. assess/critique, calculate, argue a position) so that *what is rewarded* (gets the most marks in this case) is student work that matches the verbal descriptors.

The task (examination question) is set in an everyday context yet it still requires students to demonstrate deep knowledge and understanding of Physics.

The examination question and associated marking guidelines are now presented without further comment.

Extract from NSW HSC Physics Examination

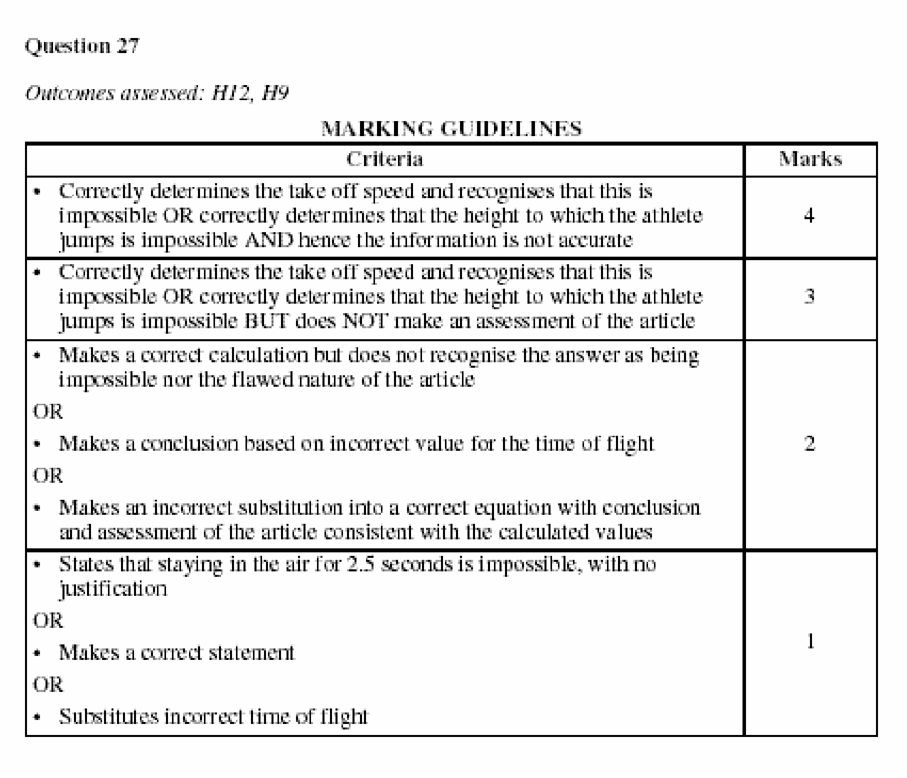
Question 27 (4 marks)

A sports magazine commenting on the athletic ability of Michael Jordan, the famous basketball player said:

‘Being an athlete takes more brains than brawn. It takes time and effort. It takes endurance and commitment. It takes an athlete who can stay in the air for 2.5 seconds while shooting a goal; an athlete who knows which laws of physics keep him there.’

Assess the information presented in this magazine, using appropriate calculations to support your argument.

[21 lines provided for written response]
[100 marks in 3 hr + 5 min reading time]



Question 27

Outcomes assessed: H12, H9

MARKING GUIDELINES

Criteria	Marks
• Correctly determines the take off speed and recognises that this is impossible OR correctly determines that the height to which the athlete jumps is impossible AND hence the information is not accurate	4
• Correctly determines the take off speed and recognises that this is impossible OR correctly determines that the height to which the athlete jumps is impossible BUT does NOT make an assessment of the article	3
• Makes a correct calculation but does not recognise the answer as being impossible nor the flawed nature of the article OR • Makes a conclusion based on incorrect value for the time of flight OR • Makes an incorrect substitution into a correct equation with conclusion and assessment of the article consistent with the calculated values	2
• States that staying in the air for 2.5 seconds is impossible, with no justification OR • Makes a correct statement OR • Substitutes incorrect time of flight	1

5.8 Objectivity and subjectivity

Can scientists do/see the Gestalt? Is the sum really greater than the parts?

There is an old stereotype, which goes like this.

End-of-semester exams and teachers are in the shared staffroom marking student scripts.

English teacher: ‘I wish I were a Maths teacher. They don’t have to read mountains of work. They just have to mark things right or wrong.’

Maths teacher: ‘Well at least my marking is objective. Yours is so subjective.’

Who is wrong? Both of them are.

There is a pervasive view in the literature that links subjectivity to writing/speaking tasks/tests with analytic or holistic marking, and that links objectivity to multiple-choice and short-answer questions with computer marking.

Actually that Maths teacher’s subjectivity was on display when she set the paper in the first place; that is, when she sampled from the domain, decided on the difficulty levels of the questions on the paper, and decided that there would be an examination paper rather than some other assessment instrument.

And the English teacher can be more objective than some multiple-choice items if s/he has set a well-crafted old-fashioned essay with unambiguous cues for students and a clear marking scheme (scoring rubric). Subjectivity/objectivity does not reside in the format but in item/test construction and scoring.

> At every stage in the design and administration of any objective test constructed by a teacher, subjective judgments are involved. The teacher has to decide on the subject matter to include, the behaviours to sample, the complexity and difficulty of proposed asks, the item format, and the wording and mode of presentation. The process is objective only at the very last stage, which is deciding on the correctness of an answer. So-called objective assessment consists of a chain of subjective decisions, with one final objective link. Unfortunately the essential objectivity of the

> end point and the fact that the outcome of the final step is often expressed in numerical form (which, to many people, is the hallmark of objectivity) obscures the subjectivity inherent in all the steps leading up to it.
>
> (Sadler, 1986)

It is understandable that some science teachers are wary of holistic grading because it is often associated with so-called subjective judgment. The notion of impression might go against the grain for a scientist who has possibly read that holistic assessment is less reliable than analytic without realising that this research usually refers to large-scale marking operations where the teacher–assessor (marker) may only have to read the script once. Some science teachers might feel comfortable with the word analytic because it has connotations of clinical. Some science teachers might have heard English teachers discussing the criteria for marking extended writing – interminable arguments about flair, structure and correct spelling, punctuation and grammar – and decided that they wanted another solution.

In pure analytic marking, the teacher–assessor assigns a mark to each of the criteria. Next question: Do you add up the marks? Do you trade off across the criteria? Do you apply a given combination rule?

5.9 Five models for dealing with trade-offs

In summary, there are five models for dealing with trade-offs:

1. No trade-offs
2. Holistic-type marking
3. Specify maximum domain performance
4. Specify minimum domain performance
5. Analytic marking and Item Response Theory.

Holistic marking and analytic marking have already been discussed in detail in this paper (sometimes under the banner of scoring or grading).

There are reservations about the minimum domain performance model (acceptable minimum standards with built-in trade-offs). It can be perceived to be a negative model, collapsing distinctions that teacher–assessors can make, thus suppressing important information about student performance.

Item Response Theory is suggested as part of the method in Appendix 4 for an empirical approach to establishing the nature of the scale that underpins teacher judgments in the new science syllabuses.

6. Applying theory and critique to existing exit criteria

Examples of two ways of presenting criteria and standards in Queensland senior science syllabuses are now identified and the instructions to teacher–assessors appear in italics below the list of criteria in each case.

6.1 Minimum standards associated with exit criteria

‘Minimum standards associated with exit criteria’ are documented in the 1995 Chemistry syllabus (p. 52).

The criteria are:

1. Knowledge of subject matter (5 standards described in terms of recalling and applying knowledge in simple situations);
2. Scientific processes (5 standards described in terms of success in simple scientific process tasks);

3. Complex reasoning processes (5 standards described in terms of using complex reasoning in challenging situations);

4. Manipulative skills (2 standards described in terms of proficiency in manipulative skills).

Allowable trade-offs for slight deficiencies in the minimum standards for knowledge subject matter or scientific processes for each exit level of achievement are outlined in [s].

Adjustments to exit levels of achievement for unsatisfactory manipulative skills are outlined in [s].

The criteria and standards are to be applied to the subject matter of this syllabus, which identifies and contextualises the senior Board subject of Chemistry for Queensland schools.

And so for Chemistry 1995, there are five standards for each of three criteria and two standards for the fourth criterion.

6.2 Standards associated with exit levels of achievement

The grading rubric that was developed by syllabus writers to guide teachers in the assessment of student folios at the end of a course of study in trial-pilot senior Chemistry is presented as a criteria/standards matrix (see Section 9 for skeleton). The approach to assessment in the new syllabus is portrayed as holistic. It has five standards. For each standard there is a description of the characteristics of a student who has attained the corresponding standard.

'Standards associated with exit levels of achievement' are documented in the 2004 Trial-Pilot Chemistry syllabus (p. 35).

The criteria are:

1. Developing knowledge and understanding – 5 standards described in terms of three sub-criteria. Sub-criterion 1 is acquiring and presenting qualitative and quantitative concepts, ideas and information. Sub-criterion 2 is recognising, comparing, classifying and explaining concepts, theories and information in processes and phenomena. Sub-criterion 3 is adapting, translating and reconstructing understandings of concepts, theories and principles;

2. Applying knowledge and understanding in societal and scientific situations – 5 standards described in terms of three sub-criteria. Sub-criterion 1 is elucidating and evaluating. Sub-criterion 2 is applying algorithms and integrating. Sub-criterion 3 is generating, critically evaluating, and justifying and so on;

3. Investigating (i.e. engaging in the research process) – 5 standards, detail not included here;

4. Using techniques (i.e. demonstrating scientific techniques) – 5 standards, details not included here.

The process of arriving at a judgment of a student folio entails matching achievement as represented by the assessment information gathered in a student folio against the exit standards as described [in the table]. This allows teachers to determine the exit level of achievement that best describes the folio as a whole. The exit standards are in a format that emphasises the holistic nature of judgments. The exit criteria are implicit in the standards.

And so for Chemistry 2004, teacher–assessors have to look at four criteria (each with multiple sub-criteria), each with five standards, make an on-balance judgment, and assign a single grade for certification. The same four criteria and their corresponding sub-criteria will also be the basis for assigning SAIs; that is, in distinguishing student achievement within an achievement band so that students in a given subject-group can be ranked as the first step in the calculation of OPs.

There is no place for 'private rules' in this high-stakes set-up.

6.3 Translation and back translation

The instruction to teacher–assessors in Section 6.2 is repeated here and then translated.

How to grade

The process of arriving at a judgment of a student folio entails matching achievement as represented by the assessment information gathered in a student folio against the exit standards as described [in the table]. This allows teachers to determine the exit level of achievement that best describes the folio as a whole. The exit standards are in a format that emphasises the holistic nature of judgments. The exit criteria are implicit in the standards.

What this means in practice

1. Look at the assessment data in the student folio.
2. Look at the exit standards as described in the syllabus.
3. Match the evidence from (1) to the corresponding description of a standard in (2) for each criterion.
4. Decide on the overall standard.
5. Assign a grade (level of achievement).

It not surprising to find out that the extent to which two independent teacher–assessors assign the same grade to a given folio is generally high: After all, the teacher–assessors are matching the characteristics of student work to a commonly-applied description of standards; and there are only five levels with very few students doing nothing (VLA) so that only four levels of achievement are really used. It is quite possible that the required level of precision could be obtained without so many words. This issue is analysed in Section 9.

More important than debating the assessment technique (holistic, analytic or whatever) is having a sophisticated and shared understanding of the meaning of the assessment criteria in the syllabus. Ensuring that the meaning of the terms used to label the criteria and sub-criteria are clear is particularly important in the following scenarios: (a) the teacher has not read the syllabus; (b) the teacher is not fully in command of subject Chemistry; or (c) the teacher is a new-comer to criteria/standards-based assessment. In the Queensland syllabus under discussion here, the four criteria are expressed in very general terms.

The communication of standards is one of the fundamental challenges in a system of standards-based assessment. Section 7 spells out the challenges involved in setting and communicating standards.

7. Setting, describing and representing standards

Much of the previous discussion leads to one conclusion – the centrality of standards in the senior system. It is inarguably challenging work to set, describe and represent standards.

What follows is a check list for syllabus writers. It refers to the general subject, S.

7.1 About standards setting

1. Are the standards pertinent to the domain of S?
2. Are the standards independent of the different curriculum organisers that schools might use to structure S for teaching–learning?
3. Do the standards reflect the legitimacy of the curriculum experience of students taking S in Queensland?
4. Are the standards representative of the demands of S as opposed to encompassing the totality of S?
5. Were experts in the discipline of S involved in setting the standards?
6. Were S teachers who were able to envision student work involved in setting the standards?

7. Are the standards realistic and attainable by the range of students doing S?
8. Do five standards per criterion really represent the expected categories of student folios sufficient to enable differentiation of achievements?

7.2 About standards descriptors

1. Are the standards descriptors fresh statements and not mere replications of curriculum objectives?
2. Does the descriptor for the highest available standard give students something to aspire to?
3. Is the descriptor for the lowest available standard written in positive (not deficit) terms?
4. Are the standards descriptors written in the language of S?
5. Do the standards descriptors describe[7] typical (as opposed to threshold) achievement?
6. Do the standards descriptors cover a range of performances on a particular criterion?
7. Do the standards descriptors clearly describe the qualities of each of the performances in the range?
8. Are the standards descriptors of an appropriate grain size (coarse- or fine-grained)?
9. Are the standards descriptors written in language that is precise yet suitable for the intended audience(s)?

7.3 About representing standards

1. Are the written standards descriptors enhanced through instantiation (i.e. able to be found in real student work)?
2. Are the standards descriptors categorised into constructs[8] that serve to justify/illustrate balance and range in experienced learning?
3. Have standards as set and described been tested out with student work and revisions made on the basis of data gathered in the exercise?

7.4 General rules

1. Differentiation between standards involves two variables, element and degree, applied separately or together.
2. Differentiation between standards is a quantitative/qualitative balancing act.
3. A hierarchy of standards can be identified at a specified juncture.
4. Standards descriptors are capable of being represented in a variety of formats (chart, matrix, dimension, poles, atoms etc.).
5. Standards descriptors should be succinct and still convey meaning.
6. Standards are words on a page plus exemplars.
7. There should be no ‘squishy notions of standards and … [settling] on what we think are reasonable marks or grades’ (Sadler, 2000).

7.5 Conditions for establishing standards

Standards are established when the following three things exist.

[7] The proposed new model (Section 10.2) does not subscribe to threshold or typical.

[8] Theoretical intangible quality or trait, which allows for individual differences in that quality or trait to be measured (constructs in action are not totally discrete).

1. Descriptors of intended standards;
2. Evidence of learning (i.e. student work) that purports to meet the standard;
3. Consensus among expert judges that the evidence does indeed meet the standard.

In other words, standards are not established overnight. They are established through words, exemplars and professional dialogue.

7.6 Application of standards schema

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

8. Other ways of assigning grades

It could be argued 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. In particular, scholarly discussion about alternative ways of assigning grades seems to have died out. At the same time, there has been a significant influx of teachers from other parts of Australia and overseas. 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 – and sometimes much is lost in translation.

8.1 Sacred status in Queensland

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 the issue under investigation in this paper. It would seem to be 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 two things: One, the lack of knowledge of the alternatives (whether these be good or bad) that could be used in the process of assigning grades; and two, the muddiness that now seems to surround some if not all of the four fundamental issues about standards.

Section 7 attended to the concept of a standard, the use of standards, standards setting, and communicating standards. Section 8.2 below summarises four alternatives for assigning grades.

8.2 Four ways of assigning grades

1. Setting numerical boundaries for grades;
2. Combination rules on pre-determined criteria;
3. Characterising a student’s global achievement in a semester/year/2-year course in terms of exit criteria;
4. Properties or characteristics or qualities of a particular piece of work.

Each of these is now discussed in more detail followed by a concluding comment about its applicability to the new senior science syllabuses.

1. Setting numerical boundaries for grades

The most commonly used grading scheme throughout the world is to award grades according to predetermined numerical ranges.

Score range	Grade
90–100	A^+
80–89	A
65–79	B
50–64	C

This scheme does not grade students 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. Here the criteria are the grade boundaries. In some place at some time, the grade boundaries were set as policy.

The use of numerical boundaries, the so-called 'cut-scores', is not ridiculous in principle. For it to be educationally sound and ethically defensible, there needs to be a great deal of attention to how the marks are generated; for example, it would require attention to subject matter and skill domains, sampling of the domain, validity of assessment items, and cut-scores related to achievement in that subject.

The standards issue in the senior sciences cannot be satisfactorily resolved by setting numerical boundaries for grades. Cut scores are not part of the culture.

2. 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 (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. Here the criteria are the composition rules. In practice this often looks like an elaborate version of cut scores with mandatory minima on different components.

The standards issue in the senior sciences cannot satisfactorily be resolved by applying composition rules. Nominating the criteria still does not tell the students anything at all about the standards.

3. 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 characteristics of a student's global achievement in a course. These are the things that a variety of subject units/topics/semesters are expected to contribute to. A determined emphasis on these would have a huge impact on curriculum planning. 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.

The criteria for Chemistry, say, could include:

- Factual (or declarative) knowledge in the subject (e.g. the structure of matter);
- Knowledge and understanding of the key concepts of the discipline (e.g. periodicity of the elements);
- Knowledge of subject-specific procedures (Chemistry-specific skills such as balancing equations for redox reactions);

- Ability to integrate these knowledges and solve problems;
- Transportable skills (generic skills) acquired in this subject, shared with other subjects, and able to be used in unrehearsed or novel situations;
- Dispositions and attitudes.

The hypothetical list above is not the ultimate list and it not a recommendation. It is provided by way of exemplification. The important point here is this: Either a determined emphasis on the assessment criteria is allowed to have a huge impact on curriculum design or curriculum design has a huge effect on the assessment criteria. Take your pick. But make sure the 'message systems' (Bernstein, 1990) are orchestrated. The global objectives at the beginning of the syllabus must map onto the exit assessment criteria at the end of the syllabus (or vice versa).

The standards issue in the senior sciences cannot satisfactorily be resolved by using fuzzy descriptors such as 'many', 'several', 'few', 'adequate', 'satisfactory', 'acceptable' and so on. The art of writing descriptors is to be precise – sometimes this involves extending the vocabulary of those who are going to use the standards descriptors. Ultimately this will benefit teacher–assessor and student. It will benefit the teacher–assessor because s/he will know exactly what to look for in student work. It will benefit the student because the assessment will be more valid.

The standards issue cannot be resolved by using *–ing* words without qualifiers that capture the sense of *how well* the student is achieving in *–ing*. Also, the term *developing*, as in the criterion 'Developing knowledge and understanding', is, strictly speaking, not a construct. Rather, the intangible quality or trait that is to be assessed is 'Knowledge and understanding'.

4. Properties or characteristics or qualities of a particular piece of work

These characteristics of a particular piece of work are different in both scope and kind from the characteristics of a student's global achievement in a course of study as in (3.) above.

The criteria below might be identified for marking a single large written task:

- Relevance to task set;
- Validity of argument, including logical development;
- Organisation of the response, including clarity of expression;
- Presentation.

In the example above, it would not be sensible to have the exit criteria written in this language, the language of one instance of assessment. This type of difference is the same type as is in the new senior science syllabuses. Section 5.5 proposed the use of scoring rubrics that contain both general and task-specific components.

The standards issue in the senior sciences cannot satisfactorily be resolved by stipulating quality criteria for individual pieces of work or academic episodes (i.e. task-specific scoring rubrics). There must also be a general component.

8.3 QSA's policy on exit assessment

QSA policy states that the following six principles must be considered (together and not individually) when a school is devising an assessment program for a 2-year course of study.

- *Information is gathered through a process of continuous assessment.*
- *Balance of assessments is a balance over the course of study and not necessarily a balance over a semester or between semesters.*
- *Exit achievement levels are devised from student achievement in all areas identified in the syllabus as being mandatory.*
- *Assessment of a student's achievement is in the significant aspects of the course of study identified in the syllabus and the school's work program.*

- *Selective updating of a student's profile of achievement is undertaken over the course of study.*
- *Exit assessment is devised to provide the fullest and latest information on a student's achievement in the course of study.*

Exit assessment (i.e. assigning one of five levels of achievement) must concurrently satisfy the six principles.

The QSA policy starts with a reference to 'devising an assessment program' (this is about curriculum planning) and finishes by referring to 'exit assessment' (this is about assessment criteria). Thus curriculum planning and assessment criteria are complementary – an orchestration of the message systems as mentioned earlier in this section. This means that any perturbation in one of curriculum and assessment will influence the other. The process of describing standards must recognise this possibility.

9. Deconstruction of existing criteria/standards matrix

A matrix is a rectangular array of elements as opposed to a chart or table.

The example of an existing criteria/standards matrix is the current criteria/standards matrix on pages 35–36 of the extended trial-pilot syllabus in Chemistry (September 2004). The matrix is labelled 'Standards associated with exit levels of achievement' and has the structure below (the verbal descriptors of standards appear in the cells). The second column has been added to the table from the syllabus document to make explicit the number of sub-criteria per criterion.

Criteria		**VH**	**H**	**S**	**L**	**VL**
1. Developing						
2. Applying						
3. Investigating						
4. Using techniques						

9.1 Appraisal of existing criteria/standards matrix

Suggested minor changes

Left-hand column should be headed ‘criterion’ not ‘criteria’ in the recognised style of table headings being expressed in the singular.

The top left-hand cell should also contain the heading ‘Standard’ (on the horizontal axis as well as ‘Criteria’ on the vertical axis) as this is the label for the top row Criteria\Standard.

As the heading for the table is ‘Standards associated with exit levels of achievement’, the headings on the columns should denote standards not levels (i.e. Very High not Very High Achievement).

The heading for the new column would be ‘Sub-criterion’.

Elements of the matrix

There are FOUR criteria for judging quality of student work.

For each criterion, there are FIVE standards.

Entries in the cells of the matrix

Criteria 1 and 2 each has THREE sub-criteria, Criterion 4 has FIVE, and Criterion 5 has FOUR.

The overall result (exit level of achievement) has FIVE standards.

The exit level of achievement is derived (holistically!) from 75 (2X5X3 + 1X5X5 + 1X5X4) available descriptions (this tally includes silences).

Details of the four criteria

	Full title of criterion	Short form
1	Developing knowledge and understanding	Developing
2	Applying knowledge and conceptual understanding	Applying
3	Investigating	Investigating
4	Using techniques	Using techniques

Sub-criteria on Criterion 1

SC1	Acquire and present qualitative and quantitative concepts, ideas and information
SC2	Recognise, compare, classify and explain concepts, theories and information in processes and phenomena.
SC3	Adapt, translate and reconstruct understandings of concepts, theories and principles.

Description of student work at five available standards on Criterion 1 according to three sub-criteria

In the strips of colour-coded information below, the presence of coloured text indicates words that are restricted to the descriptor for that standard (or one other in the case of ‘complex and challenging’); the presence of coloured lines (---) indicates where words that appear in the descriptor of the standard above do not appear in this descriptor. There are different colours for each standard being described: purple for VHA; blue for HA; green for SA; and orange for LA.

Reminder: The five strips of information all relate to ONE criterion.

It is suggested that the reader scan the colour-coded information and then return to the description in the main text.

VHA

- acquires, constructs and represents qualitative and quantitative complex and challenging ideas and concepts
- --- compares, classifies and explains concepts, theories and information about processes and phenomena, in complex situations
- adapts, translates and reconstructs understandings of concepts, theories and principles.

HA

- acquire, ---, and presents qualitative and quantitative complex and challenging ideas and concepts
- --- compares, classifies and explains concepts, theories and information about processes and phenomena ---
- adapts and translates --- understandings of concepts, theories and principles.

SA

- acquire, --- and presents qualitative and quantitative --- ideas and concepts
- ---, --- classifies and explains concepts, theories and information about processes and phenomena ---
- ---, ---, ---, interprets --- concepts, theories and principles.

LA

- ---, ---, ---, recalls and presents qualitative and quantitative --- ideas and concepts
- ---, ---, ---, ---, describes concepts and information in processes and phenomena
- ----

VLA

- restates facts
- makes statements about data and information

Scanning the colour-coded information above reveals that there are a lot of words to describe five available standards on Criterion 1 according to three sub-criteria. Sadler (2003) cautions using superfluous words when writing standards.

Even within a criterion, trade-offs are required across the sub-criteria. This means that the holistic/analytic/on-balance judgment issue occurs within as well as across criteria. Would it not be possible to illustrate the differences between standards without so much repetition? The situation above appears to be an example of the phenomenon of describing standards by way of many cells and the alteration of certain key words between cells.

And there is another source of superfluous words. There is no need to use concept, idea and principle as distinct terms in the descriptors of the sub-criteria because a *concep*t is defined as a broad abstract *idea* or guiding general *principle*.

Sometimes the way standards are written requires the teacher–assessor to 'spot the difference' – as in which key word (usually an adjective) changes from one cell to another. The situation below (not drawn from QSA syllabuses) is an example of this phenomenon.

VHA	Exceptional performance indicating complete and comprehensive understanding of the subject matter; genuine mastery of relevant skills; demonstration of an extremely high level of interpretative and analytical ability and intellectual initiative; and achievement of all major and minor objectives of the subject
HA	'Excellent' performance, then the term 'very high level' instead of 'complete and comprehensive', with other changes as appropriate
SA	'Good' performance, then 'high level' and similar terms as appropriate
LA	'Satisfactory' performance, then 'adequate', 'partial' and similar terms as appropriate

The descriptors in the table above do comply with one very important general rule about setting and communicating standards. It is repeated here from Section 7.4: Differentiation between standards involves two variables, element and degree, applied together or separately. For example, the element 'performance' is common; the variation in performance is indicated by the adjectives 'exceptional', 'excellent' etc.

The check list in Section 7 was used to analyse further the existing criteria/standards matrix. The results of that analysis follow.

Appraisal of existing criteria/standards matrix

- There is a rule of thumb for naming criteria that are identified for standards-based assessment: The categories/criteria for assessment should have labels that teacher–assessors can easily remember.
- The short forms, Developing, Applying, Investigating, and Using techniques, are easily brought to mind but it is not immediately obvious what the objects are of those verbs (verbal nouns); for example, 'developing' has connotations other than[9] developing *knowledge and conceptual understanding*. Then there is the acronym 'KCU', which is fine but is not used as a label in the matrix.
- The dimensions that allow the teacher–assessor to interrogate the evidence and make judgments about the quality of student work are not discipline-specific.
- There are a lot of words to describe all the available standards. Would it not be possible to illustrate the differences between standards without so much repetition?
- Even within a criterion, trade-offs are required across the sub-criteria. This means that the holistic/analytic/on-balance judgment issue occurs within as well as across criteria.
- Developing knowledge and *conceptual* understanding means more, surely, than operating on concepts (as in the *–ing* words that precede the bullet-point elaboration of the criterion). Concepts are the broad abstractions upon which a deep understanding of a discipline is based. They give rise to scientific models and require expression in discipline-specific terms.

[9] In fact it resonates with the notion of developmental continuum, which is another problem.

- The standards descriptors are not written in the language of the domain. They are actually written in the language of generic skills (with the name of the subject appearing occasionally).
- As (approximate) replications of general objectives 2 and 3, the standards descriptors fulfil the requirement that curriculum and assessment should be aligned, but they do not fulfil the requirement that the standards descriptors should be fresh statements.
- The standards descriptors do not clearly describe the *qualities* of student work for each of the performances *categories* in the range; that is, they do not describe standards.
- The language is the language of progression along a developmental continuum. It is debatable whether the condition of getting better at something over time applies to the sciences where students do not study a sequence of topics of equal intrinsic difficulty but, rather, study things that become increasingly difficult conceptually. The preferred interpretation is of producing evidence of quality of something at snapshots in time to the end of the course (fullest and latest).
- A scale is usually presented from left to right. For example, a continuum progresses from left to right to reflect the temporal nature of the variable. Standards represent categories of performance that are defined in terms of increasing quality of student work (the categories are labelled VLA to VHA). The left to right convention might be more appropriate than the right to left convention currently observed.
- The standards descriptors are not succinct.
- Even though the standards descriptors are wordy, they do not convey meaning with clarity and precision.
- These are the ultimate example of ‘squishy standards’.
- When there are squishy standards, teachers are unable to envision student work that might meet the standards. Reliable assessment would only occur after large-scale implementation strategies, or experience over time, or a tacit understanding amongst the practitioners. The latter is not a transparent mechanism for disseminating standards.
- The differentiation between standards does not involve *element* and *degree* – which should be applied together or separately.
- The language of standards is about ‘how well’. Where are the words (qualifiers) to denote degree?
- Does this signal a change in policy about senior assessment?

9.2 Response to the appraisal

The original purpose of this investigation was to examine the applicability of the two approaches to assessment in the new senior science syllabuses, to recommend one as the way forward, and to recommend how this one approach might be implemented.

My advice, however, is that the issue of dual approaches to exit assessment that has emerged through the writing and trialling of the new syllabuses is actually symptomatic of some deeper problems. The resolution of these problems will eliminate the need for the original argument about approaches while at the same time consolidate and simplify standards-based assessment in practice.

What follows here is one possibility for dealing with the immediate problem of the criteria/standards matrix.

9.3 Reconstruction of existing matrix into a prototype standards schema

1. Re-name the four criteria thus: Knowledge & Understanding; Application; Scientific Investigation; Techniques & Procedures.

2. Replace matrix (4 rows and 5 columns) with 4 horizontal lines.
3. Refer to dimensions not criteria. Each horizontal line represents performance on a single dimension.
4. Re-order standards (or levels of achievement) to read from lowest to highest (left to right) to reflect increasing quality of performance along the dimension.
5. Express the quality of performance on this dimension in terms of the original 3 sub-criteria but do not label them as such. Use the sets of words within the descriptors for use as sign-posts along the dimension (see 9. below).
6. Reduce the total number of words to approximately one-tenth of the current amount.
7. Use a rainbow (or single-colour screen from lightest to darkest reading left to right) to represent increasing quality of performance on the dimension (horizontally on the page) and ensure that the boundaries between colours (or screens) are not clear-cut.
8. Place the rainbows, one under the other, on the page.
9. Do not necessarily subdivide all the rainbows (or variable screenings of one colour) the same way. It might be the case that not all dimensions lend themselves to five standards.
10. Add annotations at appropriate places along the dimensions where it has been shown that evidence of student performance can be produced. It might be the case that setting equal intervals between descriptions is artificial (the metaphor of changing gears comes to mind – the driver just knows when the change is needed).

See Appendix 3 for preliminary ideas about how such a diagram might look.

10. Criteria/standards matrix or something else again?

One of the reasons for employing criteria and standards is to ensure that teacher–assessors make sound judgments based on relevant criteria, rather than on gut instinct or whimsy. (Another important purpose, not discussed in this paper, is to ensure that students are aware of the way judgments are to be made about the quality of their work.)

The simplest way of presenting criteria and standards is a chart based on the work of syllabus writers and practising teachers who have been able to agree on a list of criteria for assessment, the specifications for standards of attainment against each of those criteria, and the permitted trade-offs for arriving at an overall grade.

10.1 Traditional device

The traditional way of presenting criteria and standards for judging the quality of student performance against multiple criteria, is a matrix in which:

- row (or column) headings give the available grades;
- column (or row) headings name the criteria;
- the cells provide, for each criterion, the standards descriptors for each grade.

In common practice, every cell of the matrix contains an entry and so the number of discernible differences used is the same for all criteria.

Such matrices serve as a device that helps fulfil the broad intents of a criteria/standards approach, provided the distinctions drawn are not just pass/fail or competent/ not competent. For example, these matrices help ensure that assessment is criteria-based rather than norm-based, and that standards for the various grades are explicit and transparent. In the absence of viable alternative devices, using the criteria/standards matrix has become virtually synonymous with taking a criteria-based approach.

In designing and using traditional criteria/standards matrices, syllabus writers and teacher–assessors have, however, had to grapple with the often untoward implications of certain covert assumptions built into the matrix format itself (or fostered when teacher–assessors apply that format), but which are not foundational to a criteria/standards approach. Two examples follow.

The format of the traditional matrix requires that the number of significant and discernible differences used in judging quality be the same for all criteria. This can result in syllabus writers expending effort on manufacturing distinctions in quality where real distinctions do not exist, thus obfuscating standards, biasing grades and making discussion of standards more difficult.

Traditional formats require that the quantum of achievement between adjacent standards descriptors is also the same, or thereabouts. Not only must syllabus writers compose standards descriptors for the required number of distinctions, but also they risk biasing results if their standards descriptors do not have this quantum property.

In summary, the simplicity of the matrix format can disguise real difficulties and complexities in its design and use (K. R. Gray, 2005, personal communication).

10.2 Alternative device

There are challenges for QSA in setting and disseminating standards for new subjects and for monitoring and maintaining standards in all subjects. As a way forward, it is suggested that there be a rejuvenation of the notion of exit assessment and this should find expression in a new format for exit criteria and standards so that its design by syllabus writers and its use by teacher–assessors and standards assurance officers is not only simpler but also more effective.

The re-invention should:

- remove the unnecessary, and often counter-productive, assumptions of the matrix format;
- maintain the roles matrices play in fulfilling the broad intents of a criteria/standards approach;
- remain true to the principles of standards-based assessment;
- be capable of dealing with complicated trade-offs;
- be in harmony with the experience to date of Queensland teachers in making professional judgments about standards;
- be capable of delivering the finer grained information required in assigning SAIs.

For a given subject, in arriving at an exit level of achievement, teacher–assessors judge the quality of student performance on several (typically three) criteria, before assigning an overall grade (level of achievement). The criteria emphasise ‘big things that matter’ – rich dimensions of achievement. This is essentially the same idea as in the traditional device and the new criteria/standards matrices for Chemistry, Physics and Science21 (notwithstanding separate discussion of task-based criteria versus exit-level criteria).

Positioning teacher–assessors’ judgments along a single *dimension* (as per the definition of criterion) presented as a horizontal line rather than having them allocated to discrete cells, allows the number of standards to vary from criterion to criterion, as can their relative placements. Having variation is not an essential element of this new model. Having the possibility of variation is.

Appendix 3 shows what the new format might look like. Any acceptance of this notion would need to be accompanied by further discussion on the nature of the standards descriptors, their expression, positioning and use.

In a new model, teacher–assessors should not be forced to consign a student folio to a cell when their judgment is that no cell descriptor adequately matches the folio. It should also give due recognition to the fact that the best-written generalised standards, especially of high-quality performance, can never be sharply defined or communicated with absolute precision. This allows teacher–assessors, once they have understood the dimensions and the features of student performance on those dimensions, to focus more on the quality of the work in the folio that they are assessing than on the precise meaning of the standards descriptors to which a student folio must be matched or on an artificially imposed approach to making on-balance judgments.

What Queensland teachers need is a simple structure for expressing assessment criteria and performance/achievement standards so that they are able to concentrate their energies on the quality of their assessment. It is therefore recommended that QSA replace the current criteria/standards matrix with an alternative device, called a standards schema[10][11] for use by teachers in arriving at exit levels of achievement.

10.3 Cautions in standards writing

Almost 20 years after his seminal paper (1987) on specifying and promulgating achievement standards, Sadler (2003) pointed out particular dangers in this modern era of standards writing—atomism, matrices, and arm-chairing. Instances of these dangers are, respectively, fine-grained outcomes statements; criteria/standards schema with lots of cells containing superfluous words; and wise people in leather chairs sitting back and deciding what should be rewarded in student work, either at the task level or at the overarching level for reporting results.

The proposed new model contains none of the above but it does contain features necessary to support the nature of complex, multifaceted tasks that assess multiple knowledges, understandings, skills and dispositions. Other grading models, such as impression marking, or models using detailed rubrics, analytic marking and mechanical combination rules, would be counterproductive in that their application would tend to reduce the multidimensionality of complex tasks (such as in task-based curriculum/assessment).

It is recommended that QSA not mandate holistic grading as the basis for assigning exit levels of achievement but that its features remain in usage, where applicable, in the continuous assessments that make up the assessment program.

10.4 Formative and summative assessment

Formative assessment occurs when assessment, whether formal (e.g. testing), or informal (e.g. classroom questioning), is primarily intended for, and instrumental in, helping a student attain a higher level of performance. Formative assessment occurs prior to summative assessment; its purpose is partly to guide future learning for the student.

Summative assessment occurs when assessment is designed to indicate the achievement status or level of performance attained by a student at the end of a course of study or period of time. It is geared towards reporting or certification.

Much is written about formative and summative assessment in the genre of comparing and contrasting them and casting them as assessment *for* learning and assessment *of* learning respectively. Much of what is written presents a false dichotomy (Matters, 2006). Despite the definitions above of formative and summative assessment, it is the case that, in Queensland, formative assessment can count for summative purposes. This is the nature of continuous assessment as opposed to one single assessment experience at the end of a course of study.

There is no *necessary* distinction between formative and summative assessment in their content or conditions, although it must be acknowledged that VIPs in the UK do not hold this view. For them, purpose is everything in assessment. This is no doubt a consequence of their being totally immersed in national subject assessments in the compulsory years of schooling (which is not the case in Australia).

In my opinion, all assessment is assessment ‘of’ learning. Assessment results may be used ‘for’ a variety of purposes. The most productive of these purposes is the promotion of further learning but it

[10] This term was introduced by McMeniman in 1986. Although the recommended format is not the same as McMeniman’s, it is not in conflict with her thinking at that time.

[11] The name standards schema appropriately complements the description of the system as standards-based.

does not follow that reporting and certification are counterproductive in promoting learning, given for example, accompanying factors such as achieving motive and academic self-concept.

Effective assessment encompasses the dual and related goals of assessment of learning and assessment for learning. These are not assessments that can develop independently of each other. They both focus on improving learning, particularly those deeper forms that give rise to active and continued interest in learning.

It is therefore recommended that QSA not distinguish between these two terms in the syllabus documents. Queensland teachers have long been able to set assessment tasks that fulfil these dual purposes.

10.5 Task-based curriculum/assessment

In much of Australia there are now assessment regimes in the senior years that provide teachers and students with information about the standards that are to be expected from students after a course of study. In much of the world, there is now an emphasis on teacher-devised tasks and authentic assessment – a shift to tasks that, according to Shepard (1991), 'emulate the kind of process-based higher-order tasks thought to represent good practice'. The desire in Queensland for performance assessment sits alongside the need for standards-based assessment.

The standards schema proposed in Section 10 is true to the traditional approach to criteria/standards and is eminently suitable for application when the assessment includes students doing multifaceted tasks. Also, it does not value one purpose of assessment over another (i.e. summative or formative).

By definition, authentic assessment occurs when the assessment task is real – that is, students experience the task as it could be carried out in a non-school environment, the range of response modes is broad, and the skills developed in other subject areas are enhanced. Authentic assessment involves students in the use of relevant and useful knowledge, thinking and practical skills. In its association with the work of Newman (1996), the term refers to assessment that is both valid and useful in practice; that is, it assesses accurately what it purports to assess and it provides accessible and understandable data on students and programs in what teachers are able to translate into practical curriculum and pedagogic decisions via curriculum conversations.

Many writers and researchers see task-based curriculum/assessment as an assessment instrument that fits the futures agenda: It allows for futures orientations, authenticity (real-life settings), and connection to the world beyond the classroom. The potential of task-based curriculum/assessment is well known; its limitations are less well understood. Some are re-iterated below.

Task-based assessment requires that teacher–assessors arrived at a single grade for student performance in multiple domains in a performance-based task completed over an extended period of time. Here, references to New Basics research findings (Queensland Department of Education and the Arts, 2004) should not be discounted simply because we did not recommend that New Basics be extended. There are lessons to be learnt from that research for those implementing the new science syllabuses.

One of the findings was that Rich Tasks produced work that was 'as rich as or richer than the best of the rest' (Queensland Department of Education and the Arts, 2004). However, another finding was that task-based curriculum/assessment is extremely demanding of teachers and of students. It requires superior planning to maintain productivity over an extended period of time and requires schools to re-think basic structures and protocols. There are challenges for teachers on a number of fronts not the least being their own basic discipline knowledge. Also, students with 'agentic' learning styles (typically boys) are likely to be outperformed by students with 'communal' learning styles (typically girls).

These and other findings may or may not be generalisable to the senior schools with uni-disciplinary tasks in a high-stakes assessment regime.

Whatever the curriculum packaging and teaching strategies for science subjects, the ultimate aim is to hook students into canonical science; and 'creolisation' (Tobin, 2006) may not deliver the canon.

There is also an issue of equity of access to the ‘glittering prizes’ because one access route is via canonical science. It is also an imperative for getting a knowledge-based society (as per the Smart State). Task-based curriculum/assessment is, for example, merely a pathway not an end in its own right.

11. Conclusions

After examining the issues presented by QSA and the other issues that emerged alongside them (and documented throughout this report), I am of the opinion that it is necessary to proceed with the following actions.

1. Reinstate validity and reliability as the referents for rationalising changes to assessment policy and practice.
2. Re-visit, clarify and consolidate policy and procedures regarding exit assessment.
3. Close the loop between general curriculum objectives and exit assessment criteria by reconsidering the genre of the existing standards descriptors.
4. Agree that all roads lead to a simpler approach to criteria/standards and develop such an approach.

Some significant elements of the present situation were foreshadowed in Section 2 and discussed in previous sections of this report. They include the following.

- Existence of artificial dichotomies (e.g. holistic versus analytic, summative versus formative);
- Place of standards in criteria-based assessment;
- Futures agenda (e.g. the challenges of authentic assessment);
- Notion of trading off within and across exit criteria;
- Word overload in the assessment section of the syllabuses.

If the proposed assessment model and/or the four actions described above lack of acceptance, it would be illuminating to proceed with an empirical approach to comparing the two approaches before deeming one to be ‘more applicable in the current context’. The method for two research activities is provided in Appendix 4. The activities are:

(i) Pair-wise comparisons and comparability of standards;

(ii) IRT and nature of the underlying scales.

Such recommendations respectfully go beyond what was required according to Term of Reference No. 2. The approach modelled as per Term of Reference No. 3 addresses more than what is referred to in Term of Reference No. 1.

There are fourteen operational recommendations emanating from this investigation. They have been alluded to in the main body of this report, and are collated and presented in Section 12.

12. Recommendations

Recommendation 1: That there be a consolidation of, and renewed commitment to, standards-based assessment.

Recommendation 2: That the continued exploration and employment of task-based assessment be encouraged but not mandated until its feasibility, reliability and validity have been demonstrated.

Recommendation 3: That general and task-specific scoring rubrics be used together for science courses that are task-based.

Recommendation 4: That the notion of trading off within and across exit criteria be examined in the context of current discussions about holistic and analytic grading, and re-examined in the light of past discussions about on-balance judgments.

Recommendation 5: That holistic grading not be mandated as the technique for awarding exit levels of achievement but, rather, that there be an appreciation of its strengths and limitations, and its use, where applicable, in the continuous assessments that make up the assessment program.

Recommendation 6: That the standards descriptors be re-written to refer to the features of student work rather than to the behaviours of the students.

Recommendation 7: That QSA consider the contents of the 32-question check list on setting, describing and representing standards, revise as necessary, and distribute to syllabus writers, teachers and other stakeholders.

Recommendation 8: That QSA's policy position on exit assessment be strengthened and, in particular, that the link between general objectives and exit criteria be reinforced (albeit that exit criteria should be expressed differently from general objectives).

Recommendation 9: That the standards descriptors be rewritten in response to the appraisal in this report and to other issues raised in detail in the body of this report.

Recommendation 10: That an alternative to the current criteria/standards matrix, a standards schema, be developed for use by teachers in arriving at exit levels of achievement.

Recommendation 11: That no distinction between the terms *formative* and *summative* assessment be made in syllabus documents. (Queensland teachers in the senior school have long been able to set assessment tasks that fulfil these dual purposes.)

Recommendation 12: That, when assigning SAIs, teachers be required to annotate the proposed standards schema such as to provide the school principal and QSA with the evidence underpinning decisions about relative placements of students within achievement bands.

Recommendation 13: That research activities be undertaken in QSA's Testing & Analyses Section, according to the methodology outlined in this report for (i) Pair-wise comparisons and comparability of standards; and (ii) IRT and nature of the underlying scales.

Recommendation 14 – Overarching recommendation: That validity and reliability be reinstated as the referents for rationalising changes to assessment policy and practice.

Executive summary

This section is a collection of comments, conclusions and recommendations from the main body of the report on assessment of senior science subjects in Queensland circa 2006.

The issues explored in this report were not easy to treat in isolation from each other because, to a certain extent, each one shapes the other. For the same reason, it is not easy to lay out the comments, conclusions and recommendations in any sensible sequence. Nevertheless, the summary that follows is presented in segments separated by a horizontal line. The order of presentation of the recommendations is not significant.

Terms of reference

1. To investigate the theoretical underpinnings of the two approaches to criteria-based and standards-referenced assessment in the Queensland senior science syllabuses;
2. To make recommendations as to which approach is the more applicable in the current context;
3. To model the proposed approach, and provide advice about implementation.

The recommendations respectfully go beyond what was required according to ToR No. 2.

The approach modelled as per ToR No. 3 addresses more than what is referred to in ToR No. 1.

Challenges

The current challenges for science teachers posed by the new syllabuses have been identified by QSA as follows:

- Meeting the requirement for a variety of assessment strategies;
- Using task-specific descriptive standards in the place of marks;
- Generating integrated assessment tasks (for holistic grading);
- Making holistic judgments about standards;
- Awarding exit levels of achievement for certification in line with stated policy and generating finer-grained data for calculation of tertiary entrance ranks.

Comments

Some of the criticisms in this report can be taken to apply to the old as well as the new syllabuses, and to other subjects as well as the sciences. The time is right to ask about the current state of play in assessment of senior subjects in Queensland.

The issue of dual approaches to exit assessment that emerged through the writing and trialling of the new syllabuses is actually symptomatic of some deeper problems, the resolution of which will eliminate the need for the original argument about approaches while at the same time consolidating and simplifying standards-based assessment in practice.

As well as the major issues identified by QSA, the following issues also require attention:

1. The fact that the Queensland system of criteria-based assessment developed, not so much underpinned by theory but more so as a theory-building exercise in itself;
2. The derivation over time of an infinite number of interpretations of the traditional criteria/standards matrix;
3. The apparent confusion in notions of assessment regime, assessment criterion, assessment technique, and assessment instrument;

4. An over-emphasis on the distinction between holistic and analytic, together with the absence of a convincing reason for introducing a dichotomy in the first place.

Conclusions and recommendations

There are other ways of assigning grades apart from the procedure set down in the syllabus:

1. Setting numerical boundaries for grades;
2. Applying composition rules;
3. Using fuzzy descriptors such as ‘many’, ‘several’, ‘few’, ‘adequate’, ‘satisfactory’, or ‘acceptable’;
4. Stipulating quality criteria for individual pieces of work or academic episodes.

The standards issue in Queensland cannot satisfactorily be addressed in any of these ways even though some of them are in common usage around the world.

Despite all the good intentions that no doubt accompanied the development of new syllabuses and new approaches to assessment, the situation has become too complicated and too far removed from a few simple concepts.

The philosophy, policies and procedures in the senior system refer to a system that is essentially *standards-based* (compared with standards-referenced or criteria-based).

Recommendation 1: That there be a consolidation of, and renewed commitment to, standards-based assessment.

Task-based assessment requires that teacher–assessors arrived at a single grade for student performance in multiple domains in a performance-based task completed over an extended period of time. Task-based curriculum/assessment allows for futures orientations, authenticity (real-life settings), and connections to the world beyond the classroom. The potential of task-based assessment is well known; its limitations are less well understood.

Whatever the curriculum packaging and teaching strategies for science subjects, the ultimate aim is to hook students into canonical science. Task-based curriculum/assessment is merely a pathway not an end in its own right.

Task-based curriculum/assessment is extremely demanding of teachers and of students. It requires superior planning to maintain productivity over an extended period of time and requires school to re-think basic structures and protocols. There are challenges for teachers on a number of fronts not the least being their own basic discipline knowledge. Students with agentic learning styles (typically boys) are likely to be outperformed by students with typically communal learning styles (typically girls).

The discourse surrounding the senior science syllabuses tends to demonise traditional assessment instruments such as formal tests and examinations, and glorify large integrated tasks.

Recommendation 2: That the continued exploration and employment of task-based assessment be encouraged but not mandated until its feasibility, reliability and validity have been demonstrated.

Combining results on a series of tasks within a single course of study requires careful consideration of what is being required of students and what is being rewarded in the marking scheme (scoring rubric), and how the specific and general components of the results will map onto the features of the exit levels of achievement.

Recommendation 3: That general and task-specific scoring rubrics be used together for science courses that are task-based.

The terms *impression marking* and *analytic marking* are most often used in the context of large-scale testing or marking operations where speed and reliability are paramount.

Analytic marking is where some sort of marking scheme is employed that gives guidance to the marker about what features s/he should look for and what weighting should be given to them (or the influence that each feature's mark should have on the overall mark).

Analytic scoring rubrics allow for the separate assessment of each of several criteria. Each criterion is scored on a different descriptive scale.

Holistic scoring rubrics support broader judgments concerning the quality of the process or product. In the holistic scoring rubric, the criteria are considered together on a single descriptive scale.

Impression marking refers to situations where a marker, using his/her expert judgment, assigns a mark on the basis of an overall impression of the work's worth.

If suitably interrogated, an impression marker should be able to give some account of why marks were assigned as they were. S/he will be following some sort of private marking scheme with associated weightings or internalised trade-off rules and priorities. A perceived problem with impression marking is that markers may differ considerably in their private marking schemes and so different markers could give very different grades to a given piece of student work.

Scoring rubrics can be designed to contain both general and task specific components. If the purpose of a presentation is to assess students' inquiry skills and their knowledge of the scientific topic that is being investigated, an analytic rubric could be used that contains both a general component and a task-specific component. The investigative component of the rubric may consist of a general set of criteria developed for the evaluation of inquiries; the task-specific component of the rubric may contain a set of criteria developed with the specific topic/phenomenon in mind.

The appropriateness of the scoring rubric as an assessment *technique* depends on the purpose of the assessment. The distinction is essentially between publicly and privately agreed ways of trading off. In my opinion, it is not necessary to mandate holistic grading (or analytic) because it is seen to be the fairest way possible or the most expedient so long as the procedures are true to principles of standards-based assessment and trading off within that assessment model.

In the new senior science syllabuses, the criteria are not considered together on a single scale even though the process is referred to as holistic grading.

Holistic grading and standards-based assessment are not complementary. Holistic grading and criteria-based assessment without an elaboration of standards might be complementary.

Combining judgments on separate criteria involves trading off. There are several methods (including analytic grading) for dealing with trade-offs.

Only one piece of information per student per subject (i.e. one grade expressed as a level of achievement) is captured for certification. So, to what extent will trade-offs be allowed and how will trade-offs between differential criteria be facilitated?

Recommendation 4: That the notion of trading off within and across exit criteria be examined in the context of current discussions about holistic and analytic grading, and re-examined in the light of past discussions about on-balance judgments.

Recommendation 5: That holistic grading not be mandated as the technique for awarding exit levels of achievement but, rather, that there be an appreciation of its strengths and limitations, and its use, where applicable, in the continuous assessments that make up the assessment program.

For the student, the highest standard is a goal to aim for. For the teacher, standards are used in assessing or describing the quality of student performance.

A standards descriptor is 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. This could operate at the domain level or the task level or the examination level or on exit from a course of study. The teacher–assessor judges which one of several standards descriptors best matches the characteristics of a student's performance (or the other way around).

In Queensland, the present standards descriptors are more like grade specifications.

Recommendation 6: That the standards descriptors be re-written to refer to the features of student work rather than to the behaviours of the students.

School-based assessment requires good work programs and good assessment techniques because criteria/standards-based assessment does not adjust for variability in assessment techniques and instruments. It is concerned with the criteria that can be used for judging the quality of student work and with explicit statements about standards. For this reason, students need to be 'let into the secret' about what standards they should aspire to. This is made difficult by statements of standards in syllabuses that are not clear and specific

By having a description of the characteristics of student work within each standard category, the likelihood that two independent teacher–assessors would assign the same grade to a given folio is increased. The required level of agreement could be obtained without so many words.

It is more important to elaborate on what is meant by the exit assessment criteria than to debate the decision-making model. And it is even more important to have a common understanding of what those four very generally expressed things mean for each of the following scenarios: (a) the teacher has not read the syllabus; or (b) the teacher is not fully in command of the science subject; or (c) the teacher is a new-comer to criteria/standards-based assessment in Queensland.

Setting standards and communicating standards are two big challenges for QSA in the assessment of senior science subjects.

A check list has been compiled that contains 32 questions in six sets (about standards setting, standards descriptors, representing standards, general rules, conditions for establishing standards, and applications of a standards schema).

Recommendation 7: That QSA consider the contents of the 32-question check list on setting, describing and presenting standard, revise as necessary, and distribute to syllabus writers, teachers and other stakeholders.

QSA's policy on exit assessment

The following six principles must be considered (together and not individually) when a school is devising an assessment program for a 2-year course of study.

- *Information is gathered through a process of continuous assessment.*
- *Balance of assessments is a balance over the course of study and not necessarily a balance over a semester or between semesters.*
- *Exit achievement levels are devised from student achievement in all areas identified in the syllabus as being mandatory.*
- *Assessment of a student's achievement is in the significant aspects of the course of study identified in the syllabus and the school's work program.*
- *Selective updating of a student's profile of achievement is undertaken over the course of study.*
- *Exit assessment is devised to provide the fullest and latest information on a student's achievement in the course of study.*

Exit assessment (i.e. assigning one of five levels of achievement) must concurrently satisfy the six principles.

QSA policy starts by referring to 'devising an assessment program' (this is about curriculum planning) and finishes by referring to 'exit assessment' (this is about assessment criteria). Thus the policy fulfils the requirement that curriculum planning and assessment criteria are complementary.

Recommendation 8: That QSA's policy position on exit assessment be strengthened and, in particular, that the link between the general objectives and exit criteria be reinforced (albeit that exit criteria should be expressed differently from general objectives).

Appraisal of existing criteria/standards matrix

» There is a rule of thumb for naming criteria that are identified for standards-based assessment: The categories/criteria for assessment should have labels that teacher–assessors can easily remember.

» The short forms, Developing, Applying, Investigating, and Using techniques, are easily brought to mind but it is not immediately obvious what the objects are of those verbs (verbal nouns); for example, 'developing' has connotations other than[12] developing *knowledge and conceptual understanding.* Then there is the acronym 'KCU', which is fine but is not used as a label in the matrix.

» The dimensions that allow the teacher–assessor to interrogate the evidence and make judgments about the quality of student work are not discipline-specific.

» There are a lot of words to describe all the available standards. Would it not be possible to illustrate the differences between standards without so much repetition?

» Even within a criterion, trade-offs are required across the sub-criteria. This means that the holistic/analytic/on-balance judgment issue occurs within as well as across criteria.

» Developing knowledge and *conceptual* understanding means more, surely, than operating on concepts (as in the *–ing* words that precede the bullet-point elaboration of the criterion). Concepts are the broad abstractions upon which a deep understanding of a discipline is based. They give rise to scientific models and require expression in discipline-specific terms.

» The standards descriptors are not written in the language of the domain. They are actually written in the language of generic skills (with the name of the subject appearing occasionally).

» As (approximate) replications of general objectives 2 and 3, the standards descriptors fulfil the requirement that curriculum and assessment should be aligned, but they do not fulfil the requirement that the standards descriptors should be fresh statements.

» The standards descriptors do not clearly describe the *qualities* of student work for each of the performances *categories* in the range; that is, they do not describe standards.

» The language is the language of progression along a developmental continuum. It is debatable whether the condition of getting better at something over time applies to the sciences where students do not study a sequence of topics of equal intrinsic difficulty but, rather, study things that become increasingly difficult conceptually. The preferred interpretation is of producing evidence of quality of something at snapshots in time to the end of the course (fullest and latest).

» A scale is usually presented from left to right. For example, a continuum progresses from left to right to reflect the temporal nature of the variable. Standards represent categories of performance that are defined in terms of increasing quality of student work (the categories are labelled VLA to VHA). The left to right convention might be more appropriate than the right to left convention currently observed.

» The standards descriptors are not succinct.

[12] In fact it resonates with the notion of developmental continuum, which is another problem.

- Even though the standards descriptors are wordy, they do not convey meaning with clarity and precision.
- These are the ultimate example of ‘squishy standards’.
- When there are squishy standards, teachers are unable to envision student work that might meet the standards. Reliable assessment would only occur after large-scale implementation strategies, or experience over time, or a tacit understanding amongst the practitioners. The latter is not a transparent mechanism for disseminating standards.
- The differentiation between standards does not involve *element* and *degree* – which should be applied together or separately.
- The language of standards is about ‘how well’. Where are the words (qualifiers) to denote degree?
- Does this signal a change in policy about senior assessment?

Recommendation 9: That the standards descriptors be re-written in response to the appraisal in this report and to other issues raised in detail in the body of this report.

In designing and using traditional criteria/standards matrices, syllabus writers and teacher–assessors have had to grapple with the often untoward implications of certain covert assumptions built into the matrix format itself (or fostered when teacher–assessors apply that format), but which are not foundational to a criteria/standards approach.

The format of the traditional matrix requires that the number of significant and discernible differences used in judging quality be the same for all criteria. This can result in syllabus writers expending effort on manufacturing distinctions in quality where real distinctions do not exist, thus obfuscating standards, biasing grades and making discussion of standards more difficult.

Traditional formats require that the quantum of achievement between adjacent standards descriptors is also the same, or thereabouts. Not only must syllabus writers compose standards descriptors for the required number of distinctions, but also they risk biasing results if their standards descriptors do not have this quantum property.

What Queensland teachers need is a simple structure for expressing assessment criteria and performance/achievement standards so that they are able to concentrate their energies on the quality of the work in the folio that they are assessing rather than on the precision of the standards statements or on some artificially imposed approach to making on-balance judgments.

Recommendation 10: That an alternative to the current criteria/standards matrix, a standards schema, be developed for use by teachers in arriving at exit levels of achievement.

The proposed new model does not contain fine-grained outcomes statements; criteria/standards schema with lots of cells containing superfluous words; or wise people sitting back and deciding what should be rewarded in student work. It does contain features necessary to support the nature of complex, multifaceted tasks that assess multiple knowledges, understandings, skills and dispositions. Other grading models, such as impression marking, or models using detailed rubrics, analytic marking and mechanical combination rules, would be counterproductive in that their application and would tend to reduce the multidimensionality of complex tasks (such as in task-based assessment).

There is no necessary distinction between formative and summative assessment in their content or conditions, although it must be acknowledged that VIPs in the UK do not hold this view. For them, purpose is everything in assessment. This is no doubt a consequence of their being totally immersed in national subject assessments in the compulsory years of schooling (which is not the case in Australia).

In my opinion, all assessment is assessment *of* learning. Assessment results may be used f*or* a variety of purposes. The most productive of these purposes is the promotion of further learning but it does not follow that reporting and certification are counterproductive in promoting learning, given such accompanying factors as achieving motive and academic self-concept.

Effective assessment encompasses the dual and related goals of assessment of learning and assessment for learning. These are not assessments that can develop independently of each other. They both focus on improving learning, particularly those deeper forms that give rise to active and continued interest in learning.

Recommendation 11: That no distinction between the terms *formative* and *summative* assessment be made in syllabus documents. (Queensland teachers in the senior school have long been able to set assessment tasks that fulfil these dual purposes.)

The coarse grain of the current standards descriptors together with holistic grading is problematic in the process of assigning SAIs used in the compilation of OPs. SAIs have rank and interval properties. SAIs are supposedly assigned after the level of achievement has been awarded.

Recommendation 12: That, when assigning SAIs, teachers be required to annotate the proposed standards schema such as to provide the school principal and QSA with the evidence underpinning decisions about relative placements of students within achievement bands.

The major recommendations may not be acceptable. If they are not, it would be illuminating to proceed on an empirical basis to compare the two approaches identified by QSA before deeming one to be ‘more applicable in the current context’. If the major recommendations are acceptable, it would still be worthwhile to obtain some data to test out whether the proposed new approach is capable of accommodating all the needs of the apparently different approaches.

Recommendation 13: That research activities be undertaken in QSA’s Testing & Analyses Section, according to the methodology outlined in Appendix 4: (I) Pair-wise comparisons and comparability of standards; and (II) IRT and nature of the underlying scales.

Recommendation 14 – Overarching recommendation: That validity and reliability be reinstated as the referents for rationalising changes to assessment policy and practice.

References

Bernstein, B. (1990). *The structuring of pedagogic discourse*. London: Routledge & Kegan Paul.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139–148.

Board of Secondary Studies. (1978). A review of school-based assessment in Queensland secondary schools (ROSBA). (Professor E. Scott, Chairman).

Brown, S. & Smith, B. (1997). *Getting to grips with assessment.* SEDA Special No. 3, Staff and Educational Development Association Publication.

Brookhart, S. M. (1999). *The art and science of classroom assessment,* ASHE–ERIC Higher Education Report, *27* (1).

Cotton, J. (1995). *The theory of assessment*. London: Kogan Page.

Hawkins, B. L. (1983). *Agency and communion: An alternative to masculinity and femininity*. Paper presented at the annual convention of the American Personnel and Guidance Association, Washington, DC.

http://spectrum.troy.edu/~alexios/EDU3371/gradingscale.htm

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83–102.

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

Matters, G. N. (2004). *Two out of five is bad.* Presentation to Anglican Schools Conference, Noosa.

Matters, G. N. (2005). *Designing assessment tasks for deep thinking*. Paper presented at Curriculum Corporation conference, Brisbane.

Matters, G. N. (2006). *Using data to support learning: students, schools, systems*. Australian Education Review No. 48. Melbourne: Australian Council for Educational Research.

Maxwell, G. S. (2001). *Are core learning outcomes standards?* Brisbane: Queensland Studies Authority.

McMeniman, M. (1986). *A standards schema* (Discussion paper 3). Brisbane: Assessment Unit, Board of Secondary School Studies.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), 229–258.

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

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.

Newmann, F. & Associates (1996). Authentic *achievement: Restructuring schools for intellectual quality*. San Francisco: Jossey-Bass.

New South Wales Board of Studies. (2004). *Higher School Certificate Examination: Physics*. Sydney: Author.

Pitman, J. A. & Dudley, R. P. (1985). *The Queensland experience of criteria-based assessment*. Paper presented at the 11th annual conference of the International Association for Educational Assessment. Oxford, England.

Pitman, J. A., O'Brien, J. E., & McCollow, J. E. (1999). *High-quality assessment: We are what we believe and do*. Paper presented at the 25th annual conference of the International Association for Educational Assessment. Bled, Slovenia.

Queensland Department of Education and the Arts. (2004). *The report of the New Basics research program*. Brisbane: Assessment & New Basics Branch.

Radford, W. M. C. (1970). *Public examinations for Queensland secondary school students: Report of the committee appointed to review the system of public examinations for Queensland secondary school students and to make recommendations for the assessment of students' achievements*. Brisbane: Department of Education.

Sadler, D. R. (1986). *Subjectivity, objectivity and teachers' qualitative judgments*. (Discussion paper 5). Brisbane: Assessment Unit, Board of Secondary School Studies.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13,* 191–209.

Sadler, D. R. (2000). *Assessment and grading: How we should rethink policy and practice*. Keynote address, Australian Universities Teaching Committee Forum, Canberra.

Sadler, D. R. (2003a). *Re-visiting* s*pecifying and promulgating achievement standards*. Paper presented to the Assessment and Reporting Framework Implementation Committee, Education Queensland, Brisbane.

Sadler, D. R. (2003b). *How criteria-based grading misses the point*. Paper presented at ETL conference, Queensland College of Art, Griffith University, Brisbane.

Shepard, L. (1991). Psychometricians' beliefs about learning. *Educational Researcher*, *20*(7), 2–16.

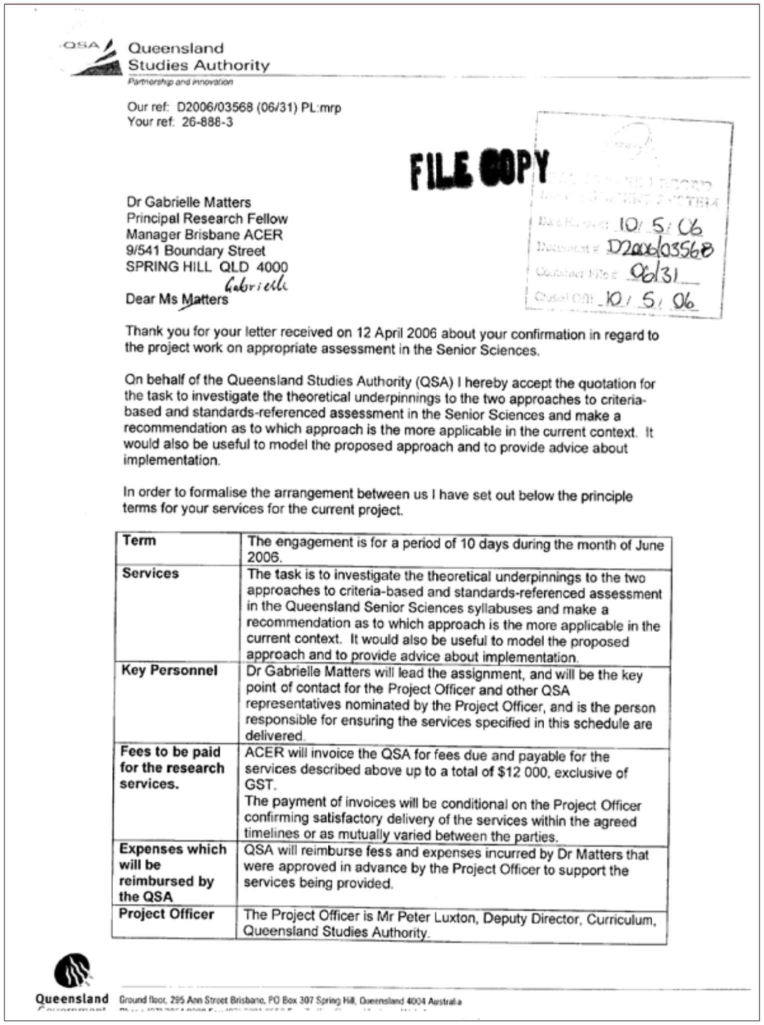
Tobin, K. (2006). *Successfully teaching and learning science in culturally diverse settings*. Paper presented at The Centre for Learning Innovation, Faculty of Education, Queensland University of Technology, Brisbane.

Tognolini, J. (2005). *Measurement*. Unpublished manuscript. Sydney: Australian Council for Educational Research.

Appendices

1. Letter of appointment
2. Letter providing background and scope
3. Working outline of prototype standards schema
4. Empirical basis for deciding on preferred approach

Appendix 1: Letter of appointment



Queensland
Studies Authority
Partnership and innovation

Our ref: D2006/03568 (06/31) PL:mrp
Your ref: 26-888-3

FILE COPY

RECORD MANAGEMENT SYSTEM
Date Received: 10/ 5/ 06
Document # D2006/03568
Container File # 06/31
Closed Off: 10/ 5/ 06

Dr Gabrielle Matters
Principal Research Fellow
Manager Brisbane ACER
9/541 Boundary Street
SPRING HILL QLD 4000

Dear ~~Ms~~ Gabrielle Matters

Thank you for your letter received on 12 April 2006 about your confirmation in regard to the project work on appropriate assessment in the Senior Sciences.

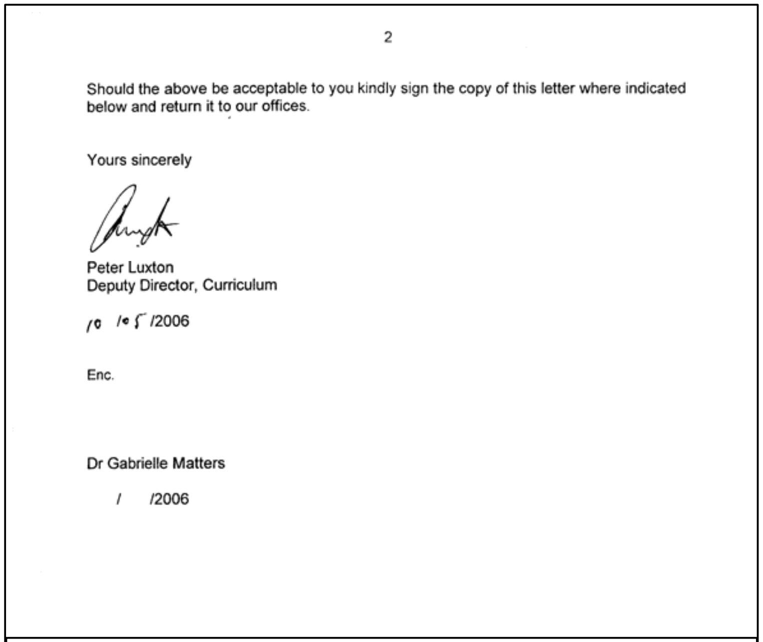
On behalf of the Queensland Studies Authority (QSA) I hereby accept the quotation for the task to investigate the theoretical underpinnings to the two approaches to criteria-based and standards-referenced assessment in the Senior Sciences and make a recommendation as to which approach is the more applicable in the current context. It would also be useful to model the proposed approach and to provide advice about implementation.

In order to formalise the arrangement between us I have set out below the principle terms for your services for the current project.

Term	The engagement is for a period of 10 days during the month of June 2006.
Services	The task is to investigate the theoretical underpinnings to the two approaches to criteria-based and standards-referenced assessment in the Queensland Senior Sciences syllabuses and make a recommendation as to which approach is the more applicable in the current context. It would also be useful to model the proposed approach and to provide advice about implementation.
Key Personnel	Dr Gabrielle Matters will lead the assignment, and will be the key point of contact for the Project Officer and other QSA representatives nominated by the Project Officer, and is the person responsible for ensuring the services specified in this schedule are delivered.
Fees to be paid for the research services.	ACER will invoice the QSA for fees due and payable for the services described above up to a total of $12 000, exclusive of GST. The payment of invoices will be conditional on the Project Officer confirming satisfactory delivery of the services within the agreed timelines or as mutually varied between the parties.
Expenses which will be reimbursed by the QSA	QSA will reimburse fess and expenses incurred by Dr Matters that were approved in advance by the Project Officer to support the services being provided.
Project Officer	The Project Officer is Mr Peter Luxton, Deputy Director, Curriculum, Queensland Studies Authority.

Queensland Government

Ground floor, 295 Ann Street Brisbane, PO Box 307 Spring Hill, Queensland 4004 Australia



2

Should the above be acceptable to you kindly sign the copy of this letter where indicated below and return it to our offices.

Yours sincerely

Peter Luxton
Deputy Director, Curriculum

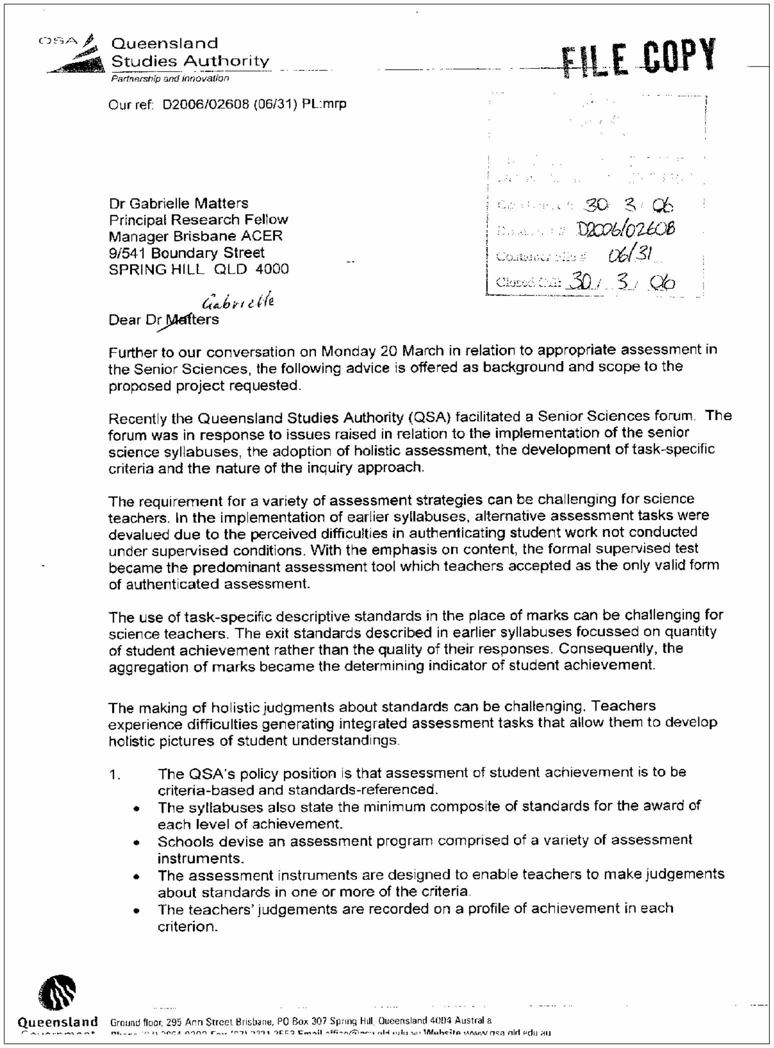
10 / 05 /2006

Enc.

Dr Gabrielle Matters

/ /2006

Appendix 2: Letter providing background and scope



QSA Queensland Studies Authority
Partnership and innovation

FILE COPY

[illegible] 30 3 06
[illegible] D2006/02608
Container File # 06/31
Closed Off: 30 / 3 / 06

Our ref: D2006/02608 (06/31) PL:mrp

Dr Gabrielle Matters
Principal Research Fellow
Manager Brisbane ACER
9/541 Boundary Street
SPRING HILL QLD 4000

Dear Dr ~~Matters~~ Gabrielle

Further to our conversation on Monday 20 March in relation to appropriate assessment in the Senior Sciences, the following advice is offered as background and scope to the proposed project requested.

Recently the Queensland Studies Authority (QSA) facilitated a Senior Sciences forum. The forum was in response to issues raised in relation to the implementation of the senior science syllabuses, the adoption of holistic assessment, the development of task-specific criteria and the nature of the inquiry approach.

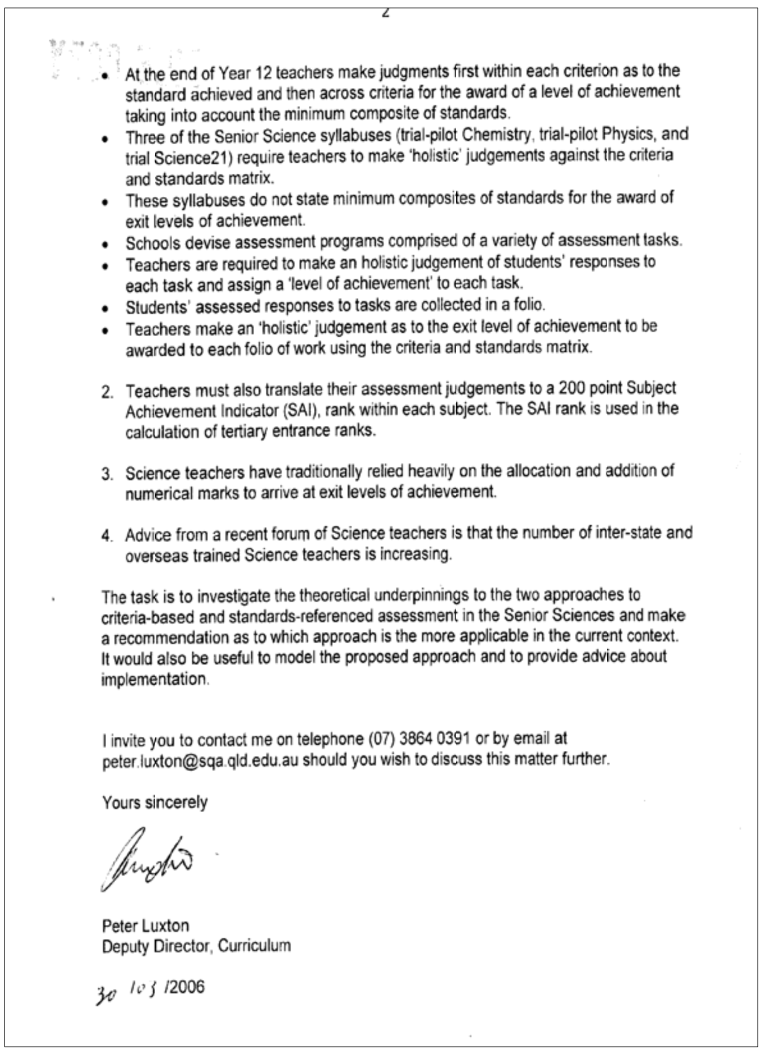
The requirement for a variety of assessment strategies can be challenging for science teachers. In the implementation of earlier syllabuses, alternative assessment tasks were devalued due to the perceived difficulties in authenticating student work not conducted under supervised conditions. With the emphasis on content, the formal supervised test became the predominant assessment tool which teachers accepted as the only valid form of authenticated assessment.

The use of task-specific descriptive standards in the place of marks can be challenging for science teachers. The exit standards described in earlier syllabuses focussed on quantity of student achievement rather than the quality of their responses. Consequently, the aggregation of marks became the determining indicator of student achievement.

The making of holistic judgments about standards can be challenging. Teachers experience difficulties generating integrated assessment tasks that allow them to develop holistic pictures of student understandings.

1. The QSA's policy position is that assessment of student achievement is to be criteria-based and standards-referenced.
   - The syllabuses also state the minimum composite of standards for the award of each level of achievement.
   - Schools devise an assessment program comprised of a variety of assessment instruments.
   - The assessment instruments are designed to enable teachers to make judgements about standards in one or more of the criteria.
   - The teachers' judgements are recorded on a profile of achievement in each criterion.

Queensland Government
Ground floor, 295 Ann Street Brisbane, PO Box 307 Spring Hill, Queensland 4004 Austral a
Phone (07) 3864 0299 Fax (07) 3221 2553 Email office@qsa.qld.edu.au Website www.qsa.qld.edu.au



2

- At the end of Year 12 teachers make judgments first within each criterion as to the standard achieved and then across criteria for the award of a level of achievement taking into account the minimum composite of standards.
- Three of the Senior Science syllabuses (trial-pilot Chemistry, trial-pilot Physics, and trial Science21) require teachers to make 'holistic' judgements against the criteria and standards matrix.
- These syllabuses do not state minimum composites of standards for the award of exit levels of achievement.
- Schools devise assessment programs comprised of a variety of assessment tasks.
- Teachers are required to make an holistic judgement of students' responses to each task and assign a 'level of achievement' to each task.
- Students' assessed responses to tasks are collected in a folio.
- Teachers make an 'holistic' judgement as to the exit level of achievement to be awarded to each folio of work using the criteria and standards matrix.

2. Teachers must also translate their assessment judgements to a 200 point Subject Achievement Indicator (SAI), rank within each subject. The SAI rank is used in the calculation of tertiary entrance ranks.

3. Science teachers have traditionally relied heavily on the allocation and addition of numerical marks to arrive at exit levels of achievement.

4. Advice from a recent forum of Science teachers is that the number of inter-state and overseas trained Science teachers is increasing.

The task is to investigate the theoretical underpinnings to the two approaches to criteria-based and standards-referenced assessment in the Senior Sciences and make a recommendation as to which approach is the more applicable in the current context. It would also be useful to model the proposed approach and to provide advice about implementation.

I invite you to contact me on telephone (07) 3864 0391 or by email at peter.luxton@sqa.qld.edu.au should you wish to discuss this matter further.

Yours sincerely

Peter Luxton
Deputy Director, Curriculum

30 /05 /2006

Appendix 3: Working outline of prototype standards schema

This appendix contains the following:

1. *Skeleton standards schema* – four dimensions, rainbow format
2. *One dimension fleshed out*, rainbow format
3. *Skeleton standards schema* – four dimensions, single-colour screen format
4. Explanatory notes

The subject used for the illustration is Chemistry. There are four exit criteria.

The one criterion used for elaboration is Knowledge & Understanding. The sub-criteria for Knowledge & Understanding are:

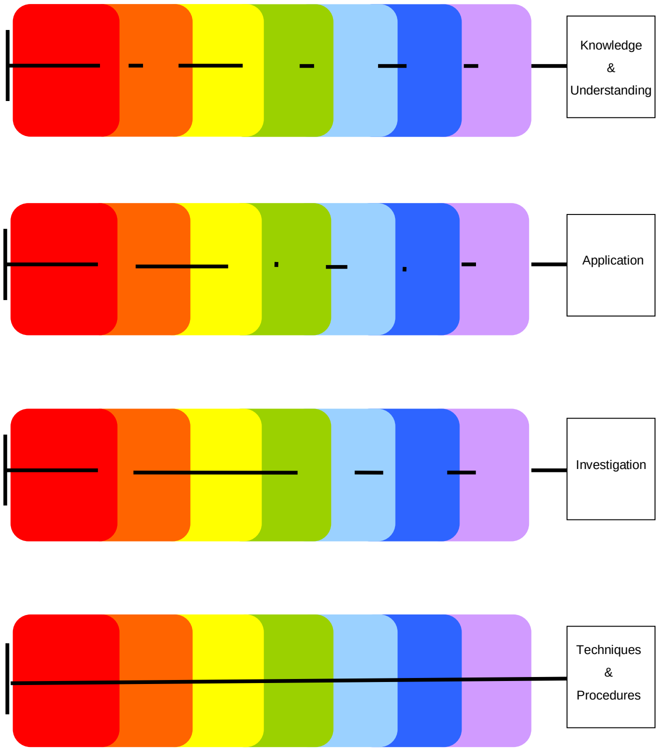
- Acquire and present qualitative and quantitative concepts, ideas and information;
- Recognise, compare, classify and explain concepts, theories and information in processes and phenomena;
- Adapt, translate and reconstruct understandings of concepts, theories and principles.

The proposed model would be a rainbow (or single-colour screen) (see 'Standards Schema' diagram on next page) with each of the dimensions detailed (see one of them, 'Dimension: Knowledge & Understanding' diagram, on next page).

Rather than a lot of words in the cells of a matrix, the dimensions model has a 'colour wash' of features of Very High Achievement in Physics to features of Very Limited Achievement in Chemistry or Physics or Science21 with interpolations along the colour axis and grade labels, not necessarily set at equal intervals along the axis, as annotations.

The information below should be studied in conjunction with Sections 9.3 and 10.2.

Standards Schema



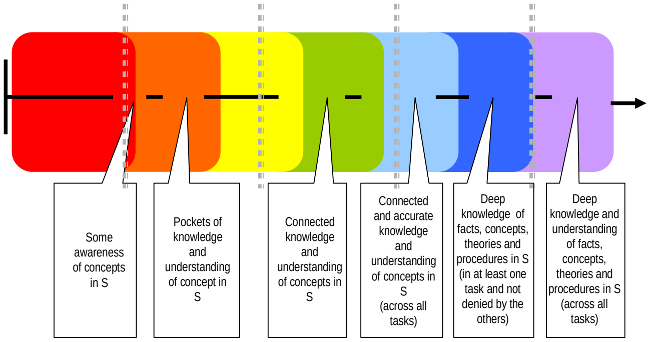
Knowledge & Understanding

Application

Investigation

Techniques & Procedures

Dimension: Knowledge & Understanding



Some awareness of concepts in S

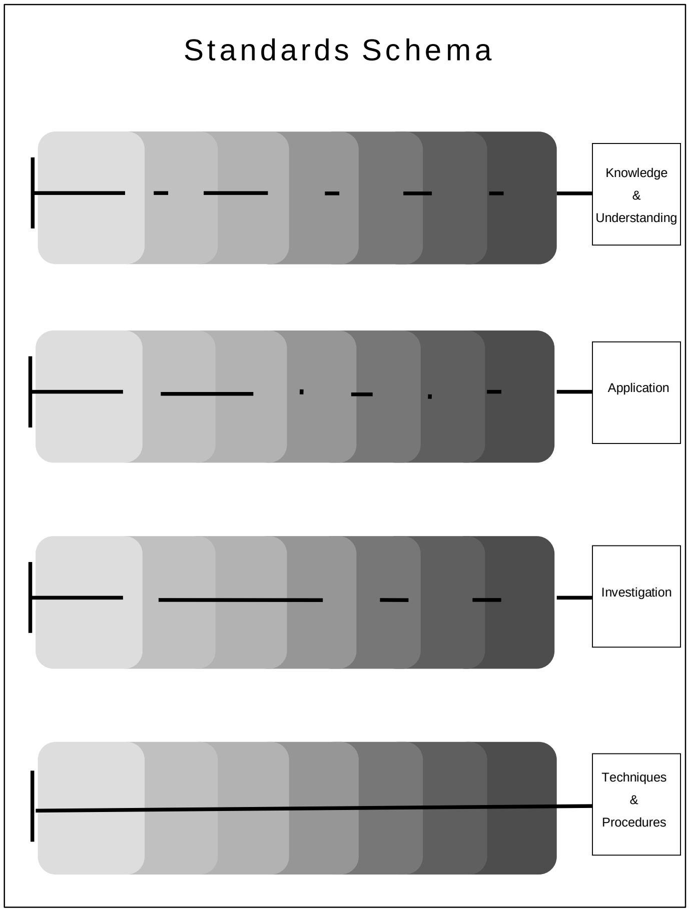
Pockets of knowledge and understanding of concept in S

Connected knowledge and understanding of concepts in S

Connected and accurate knowledge and understanding of concepts in S (across all tasks)

Deep knowledge of facts, concepts, theories and procedures in S (in at least one task and not denied by the others)

Deep knowledge and understanding of facts, concepts, theories and procedures in S (across all tasks)



Explanatory notes

- Divisions between colours of the spectrum or gradations in screening should be much fuzzier than in the diagrams presented above.
- Each dimension is labelled with full name of criterion.
- Each dimension has fine vertical lines for standards boundaries.
- It is not necessary for there always to be 5 equal divisions into standards along the sub-dimensions or dimensions. It is only necessary for there to be 5 divisions into standards for the overall assessment.
- There are annotations along each dimension expressed in the language of standards descriptors, congruent with the expression of the sub-criteria, and which discriminate, and are not necessarily situated at standards boundaries but, rather, at appropriate places where these differences are observable in student work.
- The standards schema should be accompanied by a statement about any features/properties/characteristics that cannot be traded off in making an overall judgment.
- The teacher–assessor places a cross (or coin or whatever) on each dimension at a location that matches features of student work.
- The teacher–assessor makes a balanced judgment across dimensions to arrive at final grade (level of achievement).

Appendix 4: Empirical basis for deciding on preferred approach

It is recommended that QSA undertake one or both of the activities outlined below in subjects chosen from Chemistry, Physics and Science21, as appropriate.

The description below refers to an investigation of one subject. Any comparative statements are within-subject comparisons that derive from different syllabus *versions* (e.g. old versus new) and assessment *approaches* (e.g. analytic versus holistic in new; analytic in new versus trade-off in old).

Assumptions that relate to both activities

Student learning experiences and achievement standards are not necessarily the same between syllabus versions and assessment approaches.

Research activities do not have negative effects on status of syllabus development.

Outline of Research Activity I

Assertions

Students who are awarded a VHA are 'top' students whose work exemplifies the features of a VHA as described in the exit criteria in the syllabus.

The crucial signal of what sort of work is considered to be of the highest standard is found in what students do, the work they produce under assessment conditions. It is not to be found in curriculum documents, teacher work plans or assessment instruments in isolation, no matter how well stated or set these might be.

Outline of Procedure and Analysis (using Physics as the example)

1. Acquire folios of the work of *n* very-high-achieving Physics students ('top' students) as assessed from each approach. This work would be the assessment instruments/tasks and associated marking schemes/criteria sheets plus student scripts (and not necessarily paper responses).
2. Ensure that schools/teachers do not interpret this as their standards being scrutinised but as an exemplification of excellent work from students in their schools/classes.
3. Acquire a range of judges of the quality of student work in Physics, drawn from the following groups: Physics teachers, Physics syllabus writers, and Physics assessment officers.

   The number of judges required would depend on the number of comparisons that have to be made, time available, and funding available for the exercise.
4. Ensure that there is balance of judges across categories and schools, both for validity of judgments and credibility of the process.
5. In a preliminary session, have the judges give their views about what constitutes work of top quality in Physics. Syllabus documents would be made available for this.
6. Have the judges individually examine a series of pairs of folios of student work and make individual judgments about which (if either) of each pair is of the higher standard (not 'higher standard' as in VHA rather than HA but in the general use of the term – as in 'better than …'. Ensure that identity of student/school/trial-pilot status is stripped from folios. School work programs would not be made available for this.
7. At the same time have the judges keep notes about what features of the student work contribute to decisions about its being of top quality.
8. Using the method of pair-wise comparisons (David, 1987, 1988). Some refer to this method as '(quantitative) paired comparisons of judges', which is probably more apt; some completely misunderstand and think it is simply about inter-rate agreement/correlation), arrive at rank orders of folios in terms of 'top-quality Physics'. At this stage do not equate top quality to VHA.
9. Analyse the spoken comments (from 5 above) and written comments (see 7 above) about top-quality work in senior Physics.

10. Confirm that there is a match between the student work that is at the top of the rank order list (from 8 above) and statements derived from the analysis of comments (from 9 above).
11. If there is not a match, go back to step 5 and consider the extent tot which pairs of *groups* of judges (listed in 3 above) (as opposed to pairs of individuals) agreed with each other in their judgments of the relative quality of Physics folios. This could be done in two ways: first through paired scores comparisons and second through the rank orders derived from those scores. (Alternatively, this could be a confirmatory step before 6 above.) If necessary then have an agreement trial (or training session or something) to come to a shared understanding of the meaning of top-quality work in senior Physics.
12. Match the student work from the top of the rank order list with the results of 9 (assuming that 11 has been attended to).
13. Compose standards descriptor for VHA by refining the results of 9 above.
14. Return to step 6 above to calculate the probability of a folio from a certain approach being ranked higher than a folio from the alternative approach.
15. Draw conclusions about 'top-quality Physics effect' and therefore about the comparability of standards between approaches.

What this process provides

- Confirmatory descriptionof the highest standard in Physics across Queensland;
- Comparison of student achievement across approaches; that is, an answer to the question of whether the work of students who attain a VHA under different assessment approaches is *equivalent* in terms of top-quality in Physics.

Outline of Research Activity II

Assertions

Levels of achievement that are reported on the Senior Certificate and which underpin the assigning of SAIs in the compilation of OPs are based on the same underlying scale as defined by senior subject Physics in Queensland.

An indication of uni-dimensionality and, therefore, of the effectiveness of the assessment approach in providing reliable results in the form of overall grades (VHA, HA etc.) is the 'thickness of the variable' that defines the underlying scale.

The domain definition of subject Chemistry and, to a lesser extent, subject Physics is mostly international, rather than national or local. Given that the syllabus defines the domain for the purposes of teaching and assessing the subject in Queensland, the domain definition is at least shared by Queensland educators; there is a rich pool of domain expertise to call on for research purposes.

Brief description

Information about the thickness of the variable (achievement in Physics) can be obtained by the application of Item Response Theory:

1. Obtain task-specific and student-specific data from Physics assessments in schools using the analytic approach to grading.
2. Obtain task-specific and student-specific data from Physics assessments in schools using the holistic approach to grading.
3. Analyse using Rasch modelling.
4. Draw conclusions about the underlying scale in both approaches to grading.
5. Optional, repeat using data from grading model in old syllabus and compare results.
6. Draw conclusions about the underlying scale in both versions of the syllabus.