# Review of school-based assessment (ROSBA)

Discussion papers

These papers were developed by the Assessment Unit of the Board of Secondary School Studies between 1985 and 1987.

Queensland
Board of Senior Secondary School Studies

# Contents

# ROSBA’s Family Connections

## Discussion Paper 1

**Abstract**: The Radford scheme belonged to a family of procedures known technically as “norm-referenced” assessment. The current system, called ROSBA, focuses on criteria and standards and belongs to the “criterion-referenced” family. In this Paper, something of the similarities and differences between these two families are outlined. It is also shown how ROSBA differs from the criterion-referenced testing movement in the U.S.A.

**Author**: Royce Sadler, Assessment Unit, January 1986

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



ROSBA belongs to a family of procedures known technically as “criterion-referenced assessment”. The Radford scheme, on the other hand, belonged to the “norm-referenced” family. Each family has a long and honourable history, each family can do some things better than the other. This Discussion Paper sketches in the two families as background to what ROSBA is about.

It is well known that a raw score in isolation possesses little intrinsic meaning. For instance, to say that a student has received 73% in a test does not tell us whether the level of performance is good or bad. Among other things, the test may have been exceptionally easy, or the teacher prone to award high marks for mediocre work. In some subjects, 73% would constitute failure! Similarly, to say is some context or framework that allows an index of achievement to be interpreted.

Basically, scores are given meaning by providing some basis for comparison. The three common approaches to giving scores meaning are distinguished by their comparison points. Specifically, the three approaches involve comparing a particular achievement (a) with the achievements of other students who have attempted the same tasks (“norm-referencing”), (b) with explicit descriptions of the actual things mastered (“criterion-referencing”). The last one is important in formative assessment, and in the education of the intellectually and physically handicapped. Where improvement is slow or achieved only with difficulty, any move towards success, no matter how small, is attended with celebration. Attention in this paper, however, is confined to the first two approaches because they are the ones more applicable to mainstream schooling. Note that norm-referencing and criterion-referencing represent two different approaches to interpreting and reporting student achievement, not to two different types of test items. One cannot tell just by looking at, say, a test whether it is intended to provide norm-referenced or criterion-referenced data.

## Norm-referenced assessment

Norm-referencing is the traditional method of interpreting scores. Some current (and most older) school reports record the student’s score, the class mean, and the place in class. More formally, scores may be rescaled mathematically, and converted into new or “derived” scores that make allowances for such things as the difficulty of a test, and permit more uniform interpretations. The Tertiary Entrance Score in Queensland, under both Radford and ROSBA, is an example of a norm-referenced measure. It reports the performance of a student relative to that of the total age group. Any student with a T.E. Score of, say, 975 is judged to be better academically than 97.5% of the age group, and is therefore in the top 2.5%. Technically, a T.E. Score is a (slightly modified) *centile rank*.

There are two main attractions of norm-referenced assessment. First, it provides a single score that gives a concise summary of a student’s level of achievement, even when there is no detailed analysis or inventory of particular accomplishments. Second, it simplifies decision making in situations where selections on the basis of merit must be made. This is because norm-referencing greatly facilitates ranking. But norm-referencing also has some limitations. One is that in forming aggregates of marks, the details of specific competencies are not explicitly identified at all and marks are treated as a kind of currency. A second disadvantage is that when the common practice of grading on the curve is used, the proportion of students receiving each grade is fixed, and intense competition among students often develops. Finally, an exclusive focus on norm-referenced assessment makes it difficult to monitor shifts in the performance levels of successive cohorts of students unless the test themselves remain relatively fixed.

## Criterion-referenced assessment

Criterion-referencing is an omnibus term that covers, broadly speaking, any attempt to interpret a students’s performance by referring, not to the performance of other students, but to specified domains of knowledge or behaviour. In principle, it should be possible for someone who knows a particular student’s score or achievement level to consult a catalogue that describes exactly what that student can do. Although the term “criterion-referencing” was not used until the early 1960’s, the basic idea is an old one, and motivated many widely used achievement scales in spelling, composition, drawing, handwriting, and vocabulary, dating from the turn of this century.

Part of the attractiveness of criterion-referenced assessment is the hope it holds out for a system of non-competitive assessment, in which students pit themselves against defined levels of achievement (which incorporate *standards*) rather than against one another. In doing so, of course, criterion-referenced assessment will inevitably discriminate among students, given an appropriate set of standards, and students of mixed abilities. But discrimination is not a main aim. In addition, having the nature of student achievement particularized has considerable utilitarian appeal. However in some subject areas, criterion-referencing turns out to be fairly difficult to put into practice, certainly more difficult than norm-referencing. It is therefore not surprising to find great variety in the research and development thrusts trying to find solutions to the problems. This is not the place to survey this variety, but it should be observed that the ROSBA philosophy (a) is located firmly within the criterion-referencing tradition, but (b) is sufficiently distinct from the most fully developed existing varieties in the U.S.A. for it to require independent developmental work.

Norm-referenced and criterion-referenced assessment are often portrayed as if they were mutually exclusive policies. Conceptually, it is easy (and neat) to separate them but in practice, they cannot be entirely divorced from each other, as we shall now see. Norm-referencing, in its ideal and abstract form, takes no account of the quality of student performance in an absolute sense because it is primarily concerned with orderings among students, that is, with determining which students are better than others, and by how much. Theoretically, with purely norm-referenced assessment, the average achievement level could fall year by year to some abysmally low level and no one would be any the wiser. In practice, parents and employers become concerned when students leave school after 10 or 12 years and are unable to read or write a letter, or have little appreciation of science, the culture, or the environment. This acts as a weak and indirect criterion-referenced check on what schools help students achieve. A more potent force is, of course, the integrity of teachers and the requirements set out in syllabus documents.

Similarly, a criterion-referenced system would soon come unstuck if the standards were set so absurdly low that all students were classified as “excellent”, or if standards were set so high that no one ever achieved them. Educators do have notions of what can *reasonably* be attained by students. Standards specifications that are designed by subject matter experts should reflect the range of achievements expected at a particular level of schooling. However, the specifications must not nominate in advance the proportions of students to be given the different grades. In practical terms therefore, standard setters draw on what they know students can achieve (the “norms”). Once standards are defined, the norms become irrelevant.

## Criteria and standards distinguished

As terms, standards and criteria are often used interchangeably, both in ordinary conversation and in discussions about assessment. However, a distinction can be made. Not only does it have some (but not universal) backing in the educational literature, it also turns out to be a very useful one in that it breaks the process of teacher judgment into two stages. First, the criteria have to be identified, then standards on the various criteria specified.

A *criterion* (plural: *criteria*) is a property, dimension, or characteristic by which something is judged or appraised. For example, originality, neatness, and accuracy are three criteria that could be used in assessing student projects. A *standard* is a fixed reference point for use in assessing or describing the quality of something, or as a goal to aim for. The distinction between a criterion and a standard can be clarified by means of an example, in this case one from outside education. Suppose that in testing a bicycle helmet for safety the authorities stipulate that its impact resistance must be a least 75 units for the helmet to be given A-Grade rating. Impact resistance is the criterion; 75 units is the standard, or the minimum level to be satisfied for an A-Grade rating. A helmet that tests, say, 80 units obviously meets the standard for impact resistance. Other criteria might also be important, such as weight, and visibility. Each of these criteria will have associated standards for an A-Grade helmet. B-Grade helmets would use the same criteria, but the standards would be lower. The combination of the standards on the three criteria might be called collectively “the overall standard”. Under ROSBA, the combination is called the “exit Level of Achievement”.

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 reasonable. An example of such an “absolute” standard comes from typing, where the main criteria are speed and accuracy: “The student can type at 70 words a minute with 95% accuracy”. (The difficulty of the prose, the definition of a word, and how mistakes are counted all conform to accepted conditions of testing, and are stated in the syllabus). Although there are reasons for believing that absolute standards can be devised in all areas of the curriculum, in most subjects their proper formulation will require ingenuity and persistence. This is one reason that they should probably be developed under the auspices of Subject Advisory Committees and incorporated into syllabuses. (Because they are generally difficult to construct, the temptation is often to avoid them and use norm-referencing instead.) Measurement may or may not be necessary, the essential point being that teachers will be making judgments about qualities and quality. Sometimes (as in the typing example) counting or measurement will form an essential step in determining quality or competence. In many other areas, too great a preoccupation with numbers and scores may get in the way of determinations of quality. The problem of finding workable ways to fix, define, and promulgate standards is by no means trivial, and its successful solution is likely to involve members of the Assessment Unit, Inservice Team, Review Office, and Subject Advisory Committees in extended dialogue and experimentation.

The definitions given above show that the primary focus of ROSBA is on standards rather than on criteria. However because standards presuppose criteria, an assessment system based on standards incorporates necessarily the concept of criterion-referencing. For this reason, the new scheme could be referred to generically as “standards-based assessment”.

It is necessary at this point to make crystal clear just what types of criteria form the foundation of a criterion-referenced assessment system. Without that, communication among different parties involved in education will be at cross purposes, and misunderstandings will multiply.

Criteria are the levers by which judgements are made and defended. Such judgments include classification of things or people, determinations of the most appropriate courses of action, and appraisals of quality or achievement. It is quite obvious that all of the decisions required for running and educational system employ criteria. Indeed, it is said (quite correctly) that using criteria is not particularly novel, because teachers have always had criteria in mind when setting tests. In the normal course of events, teachers have to decide which assessment instruments to use (using the criteria of objectivity, efficiency, relevance, and the like) and which items to include in a test (using such criteria as content validity, specified, difficulty level, and clarity). Other criteria are used by persons involved in the review process for judging the adequacy of Work Programs. But criterion-referenced assessment is concerned *exclusively* with the criteria that can be used for evaluating student achievements or competencies as outcomes, and *presupposes* good Work Programs and good assessment.

## Standards-based assessment and Radford

ROSBA has been variously described as a “fine tuning of” or a “radical replacement for” the Radford scheme. Both positions have some validity. ROSBA retains and strengthens many of the Radford initiatives, especially the shift to school-based curriculum development, teacher assessment of students, and a distributed responsibility for making checks and balances to the system as a whole. However, the most radical element in the new system is a shift away from norm-referenced assessment, and towards criterion-referenced principles.

One of the motivations for this change of direction (it represents more than a shift in priorities) was a determination to reduce or eliminate some of the undesirable effects that are associated with fierce competition. This change in direction will be ultimately of greater significance than the mechanics of the changeover. At present, teachers, schools, the inservice team, and review officers are more-or-less submerged in a deluge of paper, meetings, work programs, criteria and standards, assessment instruments, student folios, and some confused signals emanating from a number of sources as to what the whole exercise is about. (In addition, the requirement of tertiary institutions for selection mechanisms engages teachers in an exercise that is clearly more norm-referenced than criterion-referenced, and the duality is causing some tensions.) When the fold subsides, there will be not only a clearer conception of what standards-based assessment is, but a considerable amount will be known about how to put it into operation.

## Standards-based assessment and the criterion-referenced testing movement

Standards-based assessment is clearly a local variety of criterion-referenced assessment, and has drawn on the literature and experience elsewhere for its inspiration. In particular, it shares a concern for definitive grade specifications that set out what the various Levels of Achievement are to consist of, taking into account the various types of objectives and outcomes. But there are some points of differences. Apart from the organisational arrangements set up by the Board of Secondary School Studies to achieve comparability and consensus (while at the same time supporting school-based development), standards-based assessment is less concerned than the criterion-referenced testing movement generally with measurement as such, and with determining cut-off scores for mastery. It recognises the key role of the qualitative professional judgments of classroom teachers, both in the evaluation of individual pieces of student work, and in integrating that information to decide on a level of achievement. It puts great emphasis on the identification of criteria, the determination of standards, and the generation of policies that specify permissible trade-offs among different components. (These points are amplified in the two companion Discussion Papers “Defining Achievement Levels” and “A Standards Schema”).

This should not be taken to imply that teachers' judgments are assumed to be, *without exception*, infallible, highly reliable in the psychometric sense, or even comparable from teacher to teacher. What it does assume is that teachers by virtue of the relationship to students are in the best position to make judgments about their students' work, simply because they have access to the fullest information. But more work needs to be done in improving those judgments.

As we are finding out, the theory of a standards-based assessment is disarmingly simple but the practice is extraordinarily difficult. But having got this coveted ball into our court, we are going to see where we can hit it to maximize the good effects.

# The Case for Explicitly Stated Standards

## Discussion Paper 2

**Abstract:** Nine reasons for making criteria and standards explicit are outlined in this paper. The first six set out general benefits of being specific; the final three make a case for having explicit statements incorporated into syllabus documents.

**Author**: Royce Sadler, Assessment Unit, January 1986

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



> Why all this fuss about criteria and standards? Haven't teachers always used criteria and standards in assessing their students? Why is there this sudden advocacy for criteria and standards to be made *explicit*?

It is perfectly true that most teachers who assess the quality of student work use criteria and standards. Some teachers in assessing certain types of work have a specified set of criteria, and use it quite deliberately. Others use criteria subconsciously but could, on reflection, list the criteria used *after* having made some assessments. The ease with which this can be done depends partly on the teacher, and partly on how the subject or the task is constituted. In some subjects, for example, it is difficult for the criteria to be separated clearly from one another, because the criteria used both overlap conceptually, and are never used in isolation.

Standards are somewhat different. When teachers consistently agree in their judgments, they are probably using similar standards, even if the standards are not clearly defined. Teachers mostly carry standards around in their heads in an unarticulated form, but are still able to apply them when necessary. There are, however, a number of strong reasons for attempting to enunciate both criteria and standards. The first six in the list below refer to the general benefits of making criteria and standards explicit. The last three are about the desirability of having them stated explicitly in the syllabus or work programs, so that some measure of interschool comparability can be achieved.

### 1. Explicitness makes for clearer thinking

We often *think* we have something clear, but when we come to explain it to someone without our experience, what we thought we had clear is not so clear after all. Keeping the notions of criteria and standards distinct, and being able to derive and use them in our subject areas, is made much easier if we have some appropriate conceptual tools and the words to describe them. There is nothing mysterious about this. It is simply a manifestation of the facilitating effect that precise and appropriate language has upon thought.

### 2. Criteria and standards function more objectively

By making criteria and standards external, and not locked away in teachers' heads, some degree of objectivity is achieved. This has the potential for removing much of the impression of arbitrariness that is sometimes associated with assessment, thereby making the educational environment more open. Although achieving consensus on criteria and standards requires working at (and is sometimes difficult), articulated criteria and standards have the advantage of at making clear what the discussions should be about.

## 3. Criteria and standards help in justifying student assessments to parents

If the criteria and standards are stated clearly, the teacher can show how they are (or have been) used to appraise the quality of students' work. Properly executed, this process helps teachers defend their judgments, and can be useful in defusing accusations that the teacher is biased or shows favouritism.

## 4. Goals are clearer and appraisals less puzzling for students

Removing mystique in important if students are to understand the reasons for particular assessments, and then take an intelligent interest in improving their work. Ultimately, it will help students to become more self-evaluative. If students are unsure what they are aiming for and how their work will be judged, they are unable to exercise control over some key aspects of their learning.

## 5. Students are "judged" more by the standards than by the teacher

To the extent that standards can be externalized, they take on an existence of their own, independent of both students and teachers. *Both* parties can therefore look at them together, and co-operate in trying to achieve them. Because there are no predetermined proportions of students that can attain a given standard, one might expect an improvement in teacher-student relations.

## 6. Explicit standards help compensate for inexperience

The literature on teachers' qualitative judgments indicates that the reliability of assessments is a direct reflection of the experience of the teachers. The more experience teachers have, the more reliable their judgments generally become. Reliability here includes the agreement among different teachers on the same performances or pieces of work, and the consistency of a single teacher over time. Giving inexperienced teachers explicitly defined criteria and standards is one way of helping to bridge the experience gap.

## 7. Comparability among schools can be achieved

In deciding on whether the requirements set out in a school Work Program are appropriate for the award of the different ROSBA achievement levels, Review Panellists and Review Officers obviously have certain standards in mind. In principle, there is no reason for these not to be spelled out. (The criteria and standards referred to here apply to the assessment of students. There are, of course, quite distinct criteria that apply to the quality of the Work Programs themselves, but a consideration of those is outside the scope of the Paper.) Once standards are articulated, the logical next step is to incorporate them into the syllabus documentation. The greater the specificity of standards, the easier it is to achieve comparability across schools. Among the recommendations made by the ROSBA committee are three that have a direct bearing on the issue of comparability, namely:

P17 Board syllabuses as well as school work programs should specify objectives for the student as well as assessment criteria:

- *when objectives for the student are developed in the syllabus, Subject Advisory Committees should be mindful of standards of achievement.*
- *achievement must be related to the general capabilities of the relevant age-group. It should be interpreted not as an absolute quantity, but rather in the light of expected performance of that age-group.*

- *levels of achievement cannot be "slotted" into discrete categories which are strictly defined in minute details. Rather, achievement must be seen as part of a continuum.*

P26 *To assist in the maintenance of State-wide achievement standards and the maintenance of such standards across time, the spirit of para.6.15 of the Radford Report should be endorsed. A policy should be adopted by using 'achievement reference tests' in Board Subjects. The sole objective of these tests should be to assist schools in determining standards of performance relative to each level of achievement in a subject. Such tests will be an invaluable aid to teachers in determining achievement standards in smaller schools. Under no circumstances should the results of achievement reference tests appear on Board of school certificates or reports. Initially, achievement reference tests might be restricted to the senior secondary school with extension downwards to Year 10 depending upon the subsequent advice of the accreditation and certification committee.*

M29 *In the senior secondary school, progressive assessment for semesters one, two and three should be determined by the respective schools, and forwarded to the appropriate district subject review panel for review and comment. Any variation in assessment proposed by the panel should be a matter between the particular district panel and the school concerned—but the school concerned should have the final determination, save that any marked deviation of a school's assessment from State standards should be noted and the appropriate State panel so advised. Should a school persist in atypical standards of assessment, the matter should be reported by the State panel to the Board accreditation and certification committee for appropriate action.*

The total 'package' or design for the ROSBA system included broad statements of assessment criteria and realistic standards (P17), information generated independently be achievement reference tests (P26), and provisions for dealing with schools who's standards deviate markedly from state standards (M29). In the system as it presently operates, one important element, namely reference testing, has for various reasons not been proceeded with. The lack of one element in the design implies either that a replacement mechanism has to be developed to achieve the desired end, or that the remaining elements, which are then place under increased pressure, need to be strengthened. One way of achieving the strengthening option is to define standards explicitly, and incorporate them into syllabus documents. Not only is this approach consistent with the notion of school-based assessment, it may also be essential for achieving and maintaining public confidence in school-based assessment as part of the machinery of accountability. In addition, the processes of accreditation and certification might be made simpler for reviewing teams, and the outcome less uncertain from the school's point of view, if criteria and standards were stated explicitly.

## 8. Uniform standards cater for the mobility of students and teachers

If the standards in a particular subject are formulated independently in each school, students and teachers who transfer from one school to another are likely to encounter difficulties in the transition. This is despite the review procedures designed to ensure that the standards set in different schools are equivalent. This problem would be largely eliminated if standards were to be fixed for the state as a whole. Contrary to what one might imagine, uniformity of standards need make no inroads into the autonomy of schools with respect to selection of subject matter (where choices are available), learning experiences, and assessment programs.

## 9. Uniform standards are cost-effective

Experience overseas has shown that setting standards in many subject areas is a difficult task. It requires insight and time to identify the criteria for assessing students in different subjects, and achieving consensus on the standards in labour-intensive if the job is to be done properly. Setting them at the state level should allow the appropriate resources to be allocated.

# A Standards Schema

## Discussion Paper 3

**Abstract:** This paper presents a model for pegging standards along different criteria or performance dimensions relevant to a particular subject area. The model stands in contrast with mastery—learning forms of criterion-referenced assessment. It shows also how standards can be combined for the award of exit Levels of Achievement.

**Author**: Marilyn McMeniman, Assessment Unit, January 1986

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



## Introduction

Among other things, ROSBA represents a major shift in the nature of assessing students in the upper and lower secondary school. Under ROSBA, teachers are expected to move from the traditional, norm-referenced system of grading a student's performance by comparing it with those of other students, to a criteria-based system of assessment. This latter method entails the identification of the kinds of things a student can do in each of a number of different subjects or areas of study. Thus, quality of an individual's performance is judged in relation to some established criterion or set of criteria, rather than in relation to the work of other students.

This shift in assessment practices from a norm-based system to a criteria-based one was a major change for all those involved in the assessment of secondary school students. Key people in the ROSBA system are the subject or discipline experts who, at the school level, decide the contents of school work programs, or who, at the State level, sit on the Subject Advisory Committees (SACs) and provide guidelines for the development of the work programs. These subject experts are linchpins in the development of criteria-based assessment under ROSBA, because the success of the system rests, to a large extent, on how well the criteria and the performance levels on each of the criteria are defined. This involves the subject experts in three major undertakings:

i the identification of criteria

ii the nomination of a set of standards on each criterion along the continuum from lowest to highest proficiency; and

iii the specification of the combinations of standards to be met before a student can be judged to have reached a particular exit Level of Achievement. This involves listing the trade-offs that are acceptable at each achievement level.

The three tasks outlined above are based on an understanding of how subject disciplines are taught and learnt at secondary school level. This in turn is based on a particular notion of how knowledge is organised, and stands in contrast with the mastery learning approach to the acquisition of knowledge. Advocates of mastery learning see the ideal curriculum as comprising hermetically-sealed units which correspond to clear developmental sequences of knowledge or behaviours, each of which has to be 'mastered' before progress to a higher level is possible. However, the conceptualisation of knowledge and learning implied by the three tasks above is distinctly different. Under this latter approach, it is claimed that while *some* learning does seem to proceed in discrete steps, many important behaviours develop concurrently. This in turn implies that the performance criteria are, generally speaking, continuously relevant throughout the whole period that expertise is being acquired, even though some capabilities may be more developed than others at any point in time, and some may remain latent or dormant for long periods of time. This seems so be in accord with the way in which many teachers perceive educational growth to occur. A close comparison of mastery learning with criteria-based assessment as detailed in the sections that follow shows that meanings for the terms ‘criterion’ and ‘standard’ cannot be the same under both conceptual regimes.

Many of the difficulties encountered under ROSBA during its first years of operation can be traced to a lack of clarity as to what constitutes a criterion, and what constitutes a standard. Because of the multiplicity of interpretations many communications have been at cross purposes. The schema outlined here and the method of promulgating workable standards are based on precise definitions of criteria and standards, and the terms are used consistently throughout.

## The Schema

Essentially, the principle is that a number of standards are set along each criterion or performance dimension, classifying the range of student performances into a number of steps on the continuum from lowest to highest proficiency. The standards represent bench marks by which teachers differentiate among individual student performances. There are no predetermined requirements as to the number of such steps. It may be that only three standards, for example, are set along one particular dimension and that another dimension may have as many as six or more standards set along it.

The third task calls for some additional explanation. It relates to the identification of the combinations of standards to be met before a student can be judged to have reached a particular Level of Achievement. It is well-known that it is a rare student indeed who will reach the very top standard on every criterion. Thus the combinations of standards that have to be satisfied for the Levels of Achievement are likely to include some notion of trading-off or compensation. For example, for the award of a VHA, a student could offset a less-than-top standard of performance on two criteria by performing at top standard on all others, and similarly for students in all other achievement categories.

The following table is a representation of these three tasks as they relate to the subject *Music* at upper secondary level in Queensland schools. This representation does not appear in any official syllabus document or school work program, and should be seen only as illustrating a feasible approach to defining criteria and standards under ROSBA, and to establishing permissible trade-offs for Levels of Achievement. To convert these draft statements into a form suitable for promulgation as the official standards in Music requires further refinement and, of course, consensus on their appropriateness. However, the schema presented here shows how it is possible to define performance levels in fairly precise terms. Given this possibility, few people would quarrel with the implementation of criteria-based assessment with well-defined Levels of Achievement: within a specified period, the individual either satisfies the standards required for a particular level, or does not. There are no limits on the actual number of students who are ‘allowed’ to reach the higher levels, and conversely, there is no stipulation that there must always be some students who are allocated to the lowest Level of Achievement.

## Table 1: Music—Exit Level End Year 12 Criterion Identification [Step (i)] and Standards Specification [Step (ii)]

(to be supplemented by exemplars of students' work)

**Lowest Proficiency** **Highest Proficiency**

| Criteria | Standard 1 | Standard 2 | Standard 3 | Standard 4 | Standard 5 | Standard 6 |
|---|---|---|---|---|---|---|
| Creative Writing Skills | Student has appropriated almost no musical styles and forms and shows little ability to make any sort of statement using these styles and forms. | Student has appropriated few musical styles and forms and only occasionally can make an imaginative statement using these styles and forms. | Student has appropriated some musical styles and forms and can sometimes make a reasonably imaginative though not always very personal statement using these styles | Student has appropriated many musical styles and forms and can make a fairly imaginative and personal statement using these styles and forms. | Student has appropriated a large number of musical styles and forms and can make an imaginative and personal statement using these styles and forms. | Student has Appropriated a maximum number of musical styles and forms and can make an original, imaginative and personal statement using these styles and forms. |
| Practical Music-Making | Student almost never displays a reasonable standard of musical performance and interpretation in either instrumental or vocal areas, and cannot sight-read, improvise or conduct. | Student rarely displays a good standard of musical performance and interpretation in either instrumental or vocal areas, can sight-read only with difficulty and conducts and improvises with | Student sometimes displays a good standard of musical performance and interpretation in either instrumental or vocal areas and can sight-read, improvise and conduct when called upon but not | Student often displays a high standard of musical performance and interpretation in both instrumental and vocal areas (though perhaps not equally in both) and can also sight-read, improvise and conduct with a reasonable | Student displays a high standard of musical performance and interpretation in both vocal and instrumental areas and can also sight-read, improvise and conduct with competence. | Student displays exceptionally sensitive standards of musical performance and interpretation in both vocal and instrumental areas and can also sight-read, improvise and conduct with flair |

| Criteria | Standard 1 | Standard 2 | Standard 3 | Standard 4 | Standard 5 | Standard 6 |
|---|---|---|---|---|---|---|
| Aural Perception | Student almost never displays any aural perception in any area and almost never completes any of the given tasks. | Student sometimes displays a fair level of aural perception but has trouble with both seen and unseen works; can only rarely perform given tasks with any | Student displays a fair level of aural perception in seen work but sometimes has difficulty with unseen; can perform given tasks but not always without effort. | Student displays a good level of aural perception in seen and most unseen work and can perform most given tasks without too | Student displays a high level of aural perception in both seen and unseen work and can perform most given tasks with ease. | Student displays an outstanding level of aural perception in both seen and unseen work and can perform all given tasks with ease. |
| Musical Knowledge | Student has an extremely limited grasp of musical knowledge in all areas (historical, theoretical and stylistic) and finds the greatest | Student has a poor grasp of musical knowledge in all areas (historical, theoretical and stylistic) and has difficulty discussing issues in these areas with any degree of | Student has a reasonable grasp of musical knowledge in some areas (historical, theoretical and stylistic) and can occasionally discuss issues in some of these areas with a degree of competence. | Student has a good grasp of musical knowledge in most areas (historical, theoretical and stylistic) and can discuss issues in some of these areas | Student has an impressive grasp of musical knowledge in all areas (historical, theoretical and stylistic) and can discuss issues in most areas quite competently. | Student has an impressive grasp of musical knowledge in all areas (historical, theoretical and stylistic) and can discuss issues in these areas with insight. |

### Table 2: Trade-offs [Step (iii)—Policy On Permissible Combinations for the Awarding of Levels of Achievement on Exit]

(All four criteria are equally weighted in this example)

| Levels of Achievement | Combinations of standards to be satisfied |
|---|---|
| VHA | At this level, a student achieves at least *Standard 6 in two areas* and at least *Standard 5 in the other two areas*. |
| HA | At this level, a student achieves at least *Standard 5 in two areas* and at least *Standard 4 in the other two areas*. |
| SA | At this level, a student achieves at least *Standard 3 in three areas* and at least *Standard 2 in one area*. |
| LA | At this level, a student achieves at least *Standard 2 in three areas* and at least *Standard 1 in one area*. |
| VLA | At this level, a student achieves at least *Standard 2 in one area* and at least *Standard 1 in three areas*. |

**Source*: Ms Lorna Collingridge, Macgregor S.H.S., Visiting Teacher, Sem. 1, 1985, Department of Education, University of Queensland.

The task confronting the subject experts is a huge one and there must be some general framework within which the developmental work will be carried out. Tables 1 and 2 suggest one possible framework and show in concrete terms how it is possible to define criteria and standards and to list the combinations of standards required at each of the ROSBA Levels of Achievement. However, there are two additional points about these tables that need to be made. First, it is essential that the verbal descriptions or definitions similar to those that appear in Table 1 be supported by carefully chosen exemplars of student work. Exemplars for the standards above are currently on file but are not included here for reasons of space. Second, the trade-offs shown in Table 2 cover every combination of outcomes. Whether it treats the different combinations fairly and reasonably is quite another matter but one which could, in principle, be decided at the SAC level after gathering information from trial use in schools.

# Defining Achievement Levels

## Discussion Paper 4

**Abstract:** Statements of the five achievement levels (VLA, LA, SA, HA, VHA) constitute an important element of school Work Programs under ROSBA. In this Paper, some of the things to avoid in writing good achievement level statements are outlined. The treatment is necessarily general, but the broad principles are applicable to all subjects in the curriculum.

**Author**: Royce Sadler, Assessment Unit, January 1986

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



Statements of the five achievement levels (VLA, LA, SA, HA, and VHA) constitute an important part of school Work Programs under ROSBA. This paper sets out some of the things to avoid and the things to emphasise in writing good achievement level statements. The treatment is necessarily general, but the broad principles can be adapted to particular subjects in the curriculum. Some of the issues in the latter part of this paper are developed in more detail in the companion Discussion Paper “A Standards Schema”.

### What is meant by an “achievement level statement”?

An achievement level statement is a concise description of what a student has to be able to do at the end of a course in order to be awarded a particular achievement level (VLA, LA, SA, HA, or VHA). Although an achievement level statement is necessarily a summary, it should still be informative. A comparison of the different achievement level statements for a subject also shows what distinguishes each level from the others. Although statements of achievement levels play an important part in the specification of standards, verbal descriptions very often need supplementation with carefully selected examples if they are to come alive and have maximum utility.

### Why have such descriptions?

Properly constructed and written, achievement level descriptions have a number of virtues: (a) they help teachers to think more concretely and precisely about the standards that students might attain, (b) they help students see and understand the nature of the goal or goals they are aiming for, and (c) they help third parties, especially parents and employers, to decide whether what a student has achieved is satisfactory for a given purpose. Achievement level descriptions are written to help in specifying standards, that is, to explain the nature and quality of what students achieve. They form a potentially accessible point of contact between the school curriculum, student performance, and the outside world. In a sense, the set of achievement level descriptions is only the tip of an iceberg. Beneath the descriptions is the supporting structure of the detailed assessment schedule devised by teachers and set out in the Work Program.

## Is writing achievement level statements a difficult exercise?

In some subject areas, statements are easier to write than in others. They are easiest to write when (a) there are only a few criteria, and (b) the standards on those criteria refer to universals that can readily be interpreted outside the school setting, even by people unfamiliar with typical school work. Consider the following examples from typing:

### Example 1

The student can type X routing business letters in Y minutes with Z% accuracy.

### Example 2

The student can type non-technical prose at W words per minute with 95% accuracy.

In each of these cases there are two criteria, namely speed and accuracy, and the standards attained can be specified by means of easily interpreted numbers. Furthermore, the measurement are simple to make. The quality of performance is determined by counting the number of words typed and the mistakes made, and by measuring time. There are comparatively few subjects where standards can be laid down so easily, and with so little complication. Mostly there are multiple criteria (for example, legibility, organization, relevance of ideas, and audience are all used in evaluating student essays), and the quality of performance is determined by the professional judgments of teachers, not by counting or direct measurement.

## Desirable qualities of achievement level statements

The most informative and useful achievement level statements (a) are couched in ordinary terms and refer (as far as possible) to contexts familiar not only to teachers but also to parents, employers, and students themselves, (b) are specific in describing actual student achievements, and (c) indicate both those aspects which are firm requirements for the award of a particular achievement level, and the extent to which some accomplishments can be substituted for others.

## Terminology

The terms used in designating the ROSBA categories contain the qualifiers "high", "sound", and "limited", and the modifier "very". None of these is absolute, but each depends on a context for its meaning. To illustrate this dependence, consider an example from everyday life. If I say that a house is "expensive", I do not ordinarily mean simply that it costs a lot of money. Another house may have a higher price tag, but I may not call it expensive. I may mean "more than a house of that quality should cost" (value-for-money meaning), or perhaps "more than I could afford" (personal-capacity-to pay meaning). In either case, some external referents fill out the meaning of the term. Of course, if a statement about an expensive house occurred in conversation, the listener would either understand the meaning intended, or ask a question to settle the ambiguity. With written statements, care has to be taken to make the meaning as clear as possible at the outset.

Statements about achievement levels are difficult for the public to interpret if the context in which they are made is only poorly known to these ultimate consumers, even if teachers themselves know the context well. For example, to say that a student has "highly developed skills" spells out a dimension of interest and importance (that is, a *criterion*) but is does not really communicate a *standard* of performance. But suppose that is was reported that a student could "research a topic in the library without assistance, interview a number of people, and write a considered-opinion piece of a quality publishable in a provincial daily", the standard is fairly clear. The reason is that such a context is widely appreciated by the public in general. Note that this particular statement makes no reference to a level of schooling (such as Year 12), nor to the proportion of student who might be expected to be able to perform the task successfully. But it is absolutely necessary that such a standard be appropriate to and achievable by students at the level of schooling concerned.

## Specifity

Achievement level statements should consist of descriptions of typical achievements in a subject at each of the ROSBA levels, VHA through VLA. In particular they should spell out the special or distinctive nature of attainments in the subject and reflect its basis in a discipline. The description should lie somewhere between broad generality (such as “the student is able to recall facts, understand principles and apply them to concrete situations”) and detailed listing of subject matter (such as “the student is able to solve quadratic equations with whole number coefficients”).

### Example 3

VHA: There is evidence of thorough and extended response to the demands of each task set. The student has gone beyond the core requirements of the Syllabus to extension work involving the refinement of personal views.

*Comment*: This description is too global, refers to elements defined only in the Work Program, and is so general that it could apply to almost any subject in the curriculum.

### Example 4

HA: The student has mastered the requirements of Sound Achievement but has provided evidence of a more complete and thorough response to the demand of each task set.

*Comment*: This HA specification refers to SA, a tempting thing to do. However, there is not much point in saying only that LA is better than VLA, SA is better than LA but not up to HA, and so on.

### Example 5

SA: The student has shown a satisfactory level of achievement in the areas described and some evidence of independent research, as well as satisfactory progress in expressing a logical point of view and in evaluating issues and their significance.

*Comment*: What “satisfactory” implies may be known to the teacher, but the standard is not accessible to the reader.

### Examples 6 and 7

VHA: The student achieves at a level generally beyond the bulk of students. SA: The student will achieve average marks in all skills.

*Comments*: The weakness of these statements is that they make reference not to actual attainments but to how students achieve *relative to other students*. Unless the reader knows what the best or the average student can do, or to what certain proportions of students typically do, no one is any the wiser. Technically, describing achievement by referring to what other students do is called *norm-referencing*.

### Example 8

The student has shown a (very high, high, satisfactory, limited, very limited) level of achievement in the areas described and (much evidence, evidence, some evidence, little evidence, no evidence) of being able to carry out independent research. (The various ROSBA levels in this specification were differentiated only by the terms shown in brackets).

*Comment*: The descriptions here are all of the same form but are essentially tautological, because instead of reducing ambiguity, they say little more than that “good” is “good”. Furthermore, they differentiate one Level of Achievement from another by simple adjectives and adverbs of degree.

### Example 9

VHA: The student will attain almost perfect marks in all skills.

### Example 10

| | |
|---|---|
| VLA | 0–29% |
| LA | 30–49% |
| SA | 50–74% |
| HA | 75–89% |
| VH | 90–100% |

*Comment*: These specifications completely miss the point. They make no attempt at description, but merely cite test score ranges for the various Levels of Achievement.

## Policy on composites

Most syllabuses organize objectives into a number of classes (content, process, and skill in many subjects; listening, speaking, reading and writing in foreign languages). Obviously there is only one route to perfect achievement, and that is to achieve perfectly on all the components. In the same way there is only one way to achieve nothing at all. But it often happens that students do well on some aspects of a course and not so well on others. An achievement level statement should take these matters into account. In particular, it should indicate (a) the extent to which an exceptional degree of performance on one type of objective can be used to offset a deficiency in one or more other areas (and hence “compensate” for them), and (b) which degrees of performance (if any) are regarded as non-negotiable minima for the award of a given achievement level. Any statement that gives guidance or sets out the rules as to what combinations are and are not allowed is termed here in “policy” on composites. There are two basic policies, *compensations* and *minimum requirements*, and a third, called here a *hybrid*, which is a mixture of the two.

Nothing in what is written below should be interpreted as advocating one or other combination policy as the ideal for any particular subject. What is being advocated is that the disciplinary structure of each subject should be examined, the requirements for performance on different dimensions decided upon, and a policy formalized and incorporated in the achievement level statements. The way both teachers and students can know the rules of the game.

### Compensations

It is often assumed that all achievement levels below the highest will be characterized by progressively smaller amounts on each of the main criteria or components. To make the point clear, consider a small-scale example. Suppose that the three criteria for judging the quality of a student essay are accuracy, neatness, and originality. The highest standard will naturally be associated with very high levels on all three. Intuitively, it is perhaps not unreasonable to expect that a performance at the next highest standard will be not quite so accurate, not quite so neat, and not quite so original, and so for the lower standards of achievement. Underlying this expectation is an assumption that the levels on the components of accuracy, neatness, and originality always vary together in concert. But as examples from real life often show, such an assumption may be simplistic or false. For example, if a particular student's work is highly original, but at the same time scruffy and inaccurate, it cannot be classified easily. The problem is not with the student, but with the specification of an achievement level which, while notionally tidy, is inappropriate for students who have mixtures of high and low achievements.

The essential idea in the compensatory policy in deciding on the final Level of Achievement (VHA, HA, SA, LA, or VLA) is that good performance in one area can be "used" to make up for poor performance elsewhere. Suppose that one student is very creative or insightful but is poor at memorizing detail, while another student is just the reverse. Their performance are certainly not identical, but are they in some sense "equivalent"? That is, could high levels of insight be considered to compensate for poor recall of detail? Conversely, could memorized detail be "traded off" against insight? In particular, could both students be designated as, say, Sound Achievers?

The number of combinations theoretically possible is largest in the middle of the range. This phenomenon is not confined to the field of student assessment. It raises its head whenever sets of imperfectly correlated subscores are added together. (Statistics buffs can relate this to the Central Limit Theorem.)

Let us see why this is so by considering a simple analogy. Suppose we toss two dice and are interested only in the total score. There is only ONE way to score 12, and that is to throw 6-6. Similarly there is only one way to score 2. But there are SIX distinct combinations that give rise to the middle score of 7, namely 6-1, 5-2, 4-3, 3-4, 2-5, and 1-6.

In pre-ROSBA days, when types of objectives were not as clearly differentiated as they have to be now, the compensatory rule applied, largely by default. Teachers added up marks from assignments, knowledge tests, laboratory practical work, and field excursions quite happily. After all, marks are marks? (Well, aren't they?) Grades were then decided according to the aggregates obtained. Under ROSBA there is the necessity to decide in a more deliberate way how achievements on different criteria should be combined.

## Minimum requirements

In some areas of the curriculum, minimum attainments may be demanded on certain types of objectives for the award of, say, an HA. In a science subject, for example, it might be decided that only a student who is able to handle specimens appropriately, or to set up laboratory apparatus safely, can be awarded High Achievement, regardless of that student's breadth and depth of knowledge in other aspects of the science. (They may, of course, be "minimum qualifying levels" for each type of objective).

The collection of minimum acceptable attainments on one criterion for the different achievements levels reflects the importance of that criterion in any composite. For example, consider a hypothetical case involving the three achievement dimensions of factual knowledge, application, and manual dexterity. If only a fairly low level of manual dexterity is essential, even for the award of High Achievement, while relatively high degrees on the other two are required, this signals that manual dexterity is less important (that is, it carries a lower weighting) in the subject and therefore in the determination of an overall award than the other two.

## Hybrid policy

The compensatory and minimum requirements policies described above are obviously different from each other, but can be put together to form a hybrid policy that contains features of both. It is possible to have minimum requirements, but at the same time allow for compensations ABOVE THESE MINIMA. In the example from the previous section, although manual dexterity is effectively given a lower weighting in the composite than knowledge and application, high performance on manual dexterity may still be used to offset lesser attainments in knowledge and application for some students, provided this contingency is allowed for in the policy statement on the award of achievement levels.

The hybrid policy is an approximate model of what teachers seem to do intuitively when they review the totality of a students' work and decide on an achievement level. Elements of both minimum acceptable performance and compensation are often, but not invariably, present.

## CONCLUSION

Achievement level statements are a useful way of clarifying and crystallizing what students have to do to be awarded VHA, VLA, or something in between. The statements should, as far as possible, AVOID:

- undefined elements
- norm referencing
- tautology
- mandatory score ranges

They should, in contrast, be characterized by:

- appropriate language
- generic description of achievements
- explicit policy on permissible tradeoffs and minimum requirements

# Subjectivity, Objectivity, and Teachers' Qualitative Judgments

## Discussion Paper 5

**Abstract**: Qualitative judgments play an essential role in the assessment of student achievements in all subjects. This Discussion Paper contains a definition of qualitative judgments, a discussion of the meanings of subjectivity and objectivity, and a statement of certain conditions that must be satisfied if qualitative judgments are to enjoy a high level of credibility.

**Author**: Royce Sadler, Assessment Unit, January 1986

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



Teachers' professional judgments, like the judgments of many other professionals are about the behaviours, conditions, characteristics, or performances, of *persons*. For teachers the persons are of course their students. In assessing student work, not only do teachers engage in a deliberate act, but they are also sensitive to the consequences both for their students and for themselves of making poor judgments.

The ROSBA scheme relies heavily on the professional judgments of teachers, as did the Radford scheme it replaced. This Discussion Paper is about those professional judgments. In particular, it contains

(a) a discussion of the characteristics of a ceratin class of professional judgments which, for the sake of convenience, are called here *qualitative,*

(b) an analysis of qualitative judgments in terms of subjectivity and objectivity, and

(c) a consideration of several conditions that make for sound qualitative judgments.

Structurally, a number of general issues are dealt with first. Towards the end of the Paper, it is shown how recent developments in school-based assessments in Queensland, using criteria and standards, fit into the broader conceptualisation. Some readers might find it advantageous to skim-read this last section first, using it as an advance organiser. Throughout the Paper, criteria and standards are distinguished in an important sense. Criteria refer to dimensions or characteristics by which appraisals are made, while standards are fixed points of reference along the criteria. Criteria can exist without standards, but not the reverse.

A measure of confidence in a judgment can be obtained by comparing one judgment with a number of other judgments, all of which are made by suitably experienced members of a profession. That is, a particular judgment can be validated according to how well it agrees with the judgments of other members of the 'guild' of professionals. In practice, it is not normally expected that consensus must be reached on every single judgment. It is often sufficient to take samples of judgments for checking, or to set up procedures by which a particular judgment could, if necessary or on appeal, be compared with those of a group of experts.

External examinations were abolished in Queensland in the early 1970s. Since that time, the professional judgments teachers make about the achievements of their students have formed the basis for summative grading and certification in secondary schools. Teachers have always, of course been making professional judgments, but until the introduction of the Radford scheme, these were mostly limited to the organisational aspects of teaching, and to providing evaluative feedback to students and parents about the educational progress of students.

## Qualitative Judgments Defined

Of special interest in school-based assessment is the particular subset of professional judgments that might be called *qualitative*. A qualitative judgments is defined for the purposes of this Paper as one made directly by a teacher, in which the teacher's brain is both the source and the instrument for the appraisal. The judgments are not reducible to a formula that can be applied by a non-expert. Here "instrument" is used in a different sense from what is meant by "test instrument", in which case it means a stimulus or task specification to which students respond. An analysis of judgments that are not qualitative in the sense just defined is important philosophically, but a discussion of this is beyond the scope of this Paper. Suffice it to say that non-qualitative judgments are not necessary quantitative in that numbers or formal measures are involved.

Although technically irrelevant to the above definition, the notion of a qualitative judgment in the context of school-based assessment carries some felicitous overtones, because teachers are engaged in assessing the quality of student performance – deciding whether what a student does is, according to accepted criteria and standards, excellent, good, or bad. But judgments about quality can, under certain conditions, also be made quantitatively and mechanically.

In general, qualitative educational judgments are characterised by some or all of the following:

(a) At least some of the criteria are 'fuzzy' rather than 'sharp'. A sharp criterion contains an essential discontinuity in which there is an abrupt transition from one state to another, such as from correct to incorrect. It is always possible in principle to determine which state applies. By contrast, a fuzzy criterion is one for which there is a gradation from one state to another. Originality, as applied to an essay, is an example of a fuzzy criterion because everything between wholly unoriginal and wholly original is possible. Qualitative judgements usually involve some fuzzy criteria.

(b) There is often no independent method of confirming, at the time when a judgment is made, whether the decision or conclusion is correct. Indeed, in some circumstances, it is meaningless to speak of "correctness" at all. To give an example of methodological independence, suppose that two lamps with potential use as street lights are be compared. One way to determine effectiveness is to ask someone to judge which provides the better illumination, either in the laboratory or in the field. A difference method of judging illumination would be to set up a appropriate electro-optical instruments. These two methods are independent because they use essentially different means for arriving at a conclusion. Having several persons instead of just one would not qualify as an independent *method*, although several persons may make judgments without reference to one another, and in that sense work independently.

(c) The performances of students differ from one another on a number of dimensions. Typically there are multiple criteria, which are used simultaneously rather than sequentially. As well as individual elements or features, the total pattern of relationships among those elements is important. Because the criteria interlock, the overall configuration is what matters. In making qualitative judgments of this type, the teacher draws inferences from multiple cues. Similarly because the raw evidence is not unidimensional, the notion of *equivalences* between different performances is a key concept.

(d) There exists a large pool of potential criteria that could legitimately be brought to bear in particular assessments, but only a relatively small subset are used at any one time. The competent judge is not only able to make an appraisal, but can also substantiate a judgment by appeal to relevant criteria. Professional judgment consists in knowing the rules for using (or occasionally breaking) the rules.

(e) If numbers are used, they are assigned *after* the qualitative judgment has been made, not the reverse. Consider the situation where an overall conclusion is arrived at by adding together numerical marks (from, say, different test items, or components in a subject) and then looking at only the sheet magnitude of the result. The compounding process in this case is mechanical and therefore, according to the definition above, not qualitative. This is so in spite of the fact that the contributing judgments that give rise to the marks may themselves be qualitative.

## Subjectivity and Objectivity

No consideration of the nature of qualitative judgment proceeds far before the matters of subjectivity and objectivity are raised. The definition of qualitative judgment given earlier in this Paper implies that ALL qualitative judgments are, in a restricted sense, subjective. This is simply because the appraising apparatus is the brain of a person (that is, a *subject*, not an *object*).

This is not the way the term subjective is popularly used. It is often used in a derogatory sense when referring to judgments that are skewed or unbalanced. That is why teachers sometimes apologise for their own subjective judgments, or tend to dismiss those of others as “mere opinion”. But of course any judgment, subjective or otherwise, that is in accord with known facts or accepted contentions is, by definition, a good one. The bad reputation of subjective judgments arises from a failure to distinguish between judgments that are based on sound evidence and reasoning, and those that are based on personal taste or whimsy. They all tend to get tarred with the one brush. Subjective judgments *can* be arbitrary, biased, or slanted. On the other hand, they can be quite dependable. What must be done is to identify and create those conditions that lead to appraisals that are potentially reproducible and independently confirmable.

The transition from subjective, meaning *person*-al, to subjective, meaning biased, probably stems from a thought sequence somewhat as follows:

(a) a judgment is subjective, by definition, if a *person* makes it;

(b) the human brain processes facts in a private, value-laden environment, the resultant judgments often being idiosyncratic; therefore

(c) personal judgments cannot, as a rule, be taken at face value.

Now it is a matter of both research and common knowledge that the brain, as information processor, is not perfect. Indeed, many systematic tendencies towards distortion have been identified in the literature on human judgmental processes. But is it a categorical mistake to conclude that because *some* subjective judgements turn out to be biased and not in accord with the facts that *all* such judgments are therefore biased, or at least suspect. If one were to apply that line of thinking to daily life, things would quickly grind to a halt. Life is simply saturated with subjective judgments, most of them being made without much conscious thought. Even the most savage critic of subjective judgments could not afford to reject them out of hand. The challenge facing educators is to devise systems in which subjective judgments are made within a context of checks and balances so that the probability of unbalanced appraisals is minimised.

In relation to student assessment, proposition (b) calls for some elaboration. The literature on subjective assessment contains many depressing accounts of the fallibility of teachers’ judgments. A number of effects are well established. These include unreliability (both inter-rater, discrepancies, and the inconsistencies of one rater over time), order effects (the carry over of positive or negative impressions from one appraisal to the next, or from one item to the next on a test paper), the halo effect (letting one’s personal impression of a student interfere with the appraisal of that student’s achievement), a general tendency towards leniency or severity on the part of an assessor, and the influence of extraneous factors (such as neatness of handwriting). The reasons for these effects are multiple: the intrinsic difficulty of maintaining similar judgments have to be made), fatigue and boredom, carelessness, capriciousness, cursoriness, and personality clashes.

All of these appear to place subjective judgments under something of a cloud. However, much of the research cited by critics of subjective assessment has been undertaken with teachers working intuitively, with minimal external control by way of rules or guidelines over the processes involved. On the other side, there is unequivocal evidence that teachers who assess within a clear assessment frame work, and who have been trained in the use of the appropriate tools, made consistently reliable and valid appraisals. There are no grounds at all for being wary of teachers' subjective assessments, provided they are carried out according to carefully worked out procedures. Teachers' qualitative judgments continue to lie at the heart of good teaching and good assessment. The issue of appropriate external controls is taken up in the next section of this Paper.

At this point, something needs to be written about objectivity. Again, two meanings may be distinguished. According to the dictionary, something is said to be objective if it is real, belongs to the external world, is observable or verifiable, and exists independently of the mind.

There are certain aspects of assessment in some subjects (such as mathematics and sciences) where objectivity is understood quite naturally in more or less its dictionary sense. There is not a shadow of doubt when chemical equations are correctly balanced, mathematical problems solved, or theorems proved. There is no need to seek confirmation from a chemist, mathematician or teaching colleague. That one aspect of assessment is objective does not, however, make assessment in, say, the sciences and mathematics generally free of subjective judgments. It is incorrect to extrapolate backwards and claim that because of the existence of intrinsic objectivity at the end point that all the processes of assessment are similarly objective and therefore superior to assessment in, say, the humanities.

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 tasks, 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. So even in areas of the curriculum traditionally associated with objective assessment, the issue of subjective judgments needs to be taken seriously.

The second meaning for objectivity is derived from the first, although it is not identical with it. For qualitative judgments, objectivity has no automatic or self-evident meaning, in that everyone recognises it as factual, immediately, once it is pointed out to them. Whatever definition is adopted becomes stipulative. That is, objectivity for a particular purpose if defined into existence. Sometimes this concept of objectivity manifests itself in a convention which, in itself, is arbitrary but which becomes normative (and even binding) through widespread acceptance and common usage.

For example, an agency which produces standardised tests typically has quite strict rules and criteria for constructing, trialling, and including items in the test. The agency many decide to reject all items that have discrimination indices below a certain value. The fact that such an unambiguous rule exists does not alter the fact that the critical cut-off value could have been set higher or lower.

The common definition of objectivity in the social sciences generally, and certainly in relation to qualitative assessment, is in terms of consensus, or inter-rater reliability. This accords somewhat with the dictionary definition, because when a number of competent judges agree on an assessment, the judgment is made observable, certifiable, and exists independently of (one) mind, or at least checked, by a group of teachers. In teaching, it is reasonable to trust a single qualitative judgment provided that the person making the judgment has been 'calibrated'. If a particular assessor's judgments are refined until they are found to agree consistently with those of a number of other competent assessors, and if by appeal to the consistency and character of the assessor there is not reason to believe that the person when it is all over, the assessment can be usually accepted with confidence. The fact that a *person* directly makes a judgment has, of itself, little baring on the matter. It should now be possible to see how a single appraisal made by a teacher (even in isolation) could be both subjective and objective at one and the same time!

## Improving Qualitative Judgments

At this point, it is convenient to think of teachers-as-qualitative-judges as operating in a dual mode.

(a) as custodians of in-the-head standards, and

(b) as experts in making complex comparisons.

In-the-head standards rely to a great extent on memory, and the degree to which the various levels of performance are internalised by the teacher. Standards internalised by experienced teachers tend to be fairly stable, and for teachers who have participated in discussions and decisions at moderation meetings, to be much the same from teacher to teacher. On the other hand, in-the-head standards for inexperienced teachers tend to decay, or to undergo metamorphosis during evaluation activity. Certain so-called ‘serial’ effects, for example, are well established in the literature. The standard that an inexperienced teacher uses for grading the first in a set of student essays is often different from the standard used towards the end. Not only may high expectation at the start be unconsciously lowered if it is discovered that very few students perform at all well (if, for instance, the student do not answer, or misinterpret, the question or nature of the task). It is this malleability of in-the-head standards that makes periodic moderation or recalibration of teacher judgments necessary for inexperienced teachers.

Given that the decision as to a student’s level of achievement should be based on the fullest information available, be relatively unaffected by small aberrations in teacher judgment or student performance, and be comparable from school to school, consider the following three conditions.

*Condition A* provides opportunity to make well-considered, grounded assessments with a maximum of reflection and evidence and a minimum of time constraints. This is achieved by spreading the decision over both time and tasks, so that no single piece or testing episode is crucial.

*Condition B* provides mechanisms by which teachers can come to a consensus as to what constitutes performances of different qualities. This is achieved by setting up procedures for consultation, multiple marking, or cross validation, to enable teachers to test their own judgments against those of their peers.

*Condition C* externalizes the standards, removing them from the private knowledge structures of teachers. This makes standards accessible not only to other teachers, but also to non-members of the guild of professionals, namely parents, employers, and most importantly, the students themselves.

Condition A is intended to reduce the discrepancy between what are referred to in the literature as *performance* and *competence.* What a student actually does or produces (and therefore what the teacher assesses) is called a performance. Competence is the name given to the underlying degree of proficiency that gives rise to particular performances. Competence cannot be assessed directly, but must be inferred from a series of performances. Reasonably strong inferences can be made by considering a multiplicity of performances. A single performance may not reflect competence accurately for a variety of reasons, among them poor task specifications, sub-optimal conditions under which the students works, and imperfect teacher judgment. Condition A is designed to minimise both random fluctuations in the judgments of individual teachers and irregularities in the performances of students.

But there is a more fundamental philosophical reason for preferring evidence from multiple performances. Competence is broader in concept than a mere ‘envelope’ of performances, and by definition can be assessed only by giving students opportunity to perform a variety of tasks, under a variety of conditions. Among other things, this provides a better simulation of production conditions in the world outside the school than can ever be achieved under, say, formal (and particularly external) examinations. School-based assessment makes it feasible to think of assessing not only performance, but competence.

Condition B is obviously intended to minimise differences among the intuitive or 'base-level' standards of different teachers, which ordinarily stand a good chance of being idiosyncratic. Condition B corresponds to the definition of objectivity as agreement among competent assessors. The safety-in-numbers principle implies that the greater the number of judges who concur, the greater the objectivity.

As for Condition C, it corresponds more to the dictionary definition of objectivity, as relating to something observable and external to the mind. *Although an educational standard is, strictly speaking, always an abstraction (and never an object), the concrete referents associated with it are explicit, and can be discussed and argued about.* The most effective way of specifying and promulgating standards is a matter beyond the scope of this paper. In most subjects, however, there are sound theoretical reasons for defining standards by a combination of both verbal descriptions and carefully selected examples (called *exemplars*).

Earlier it was suggested that in many contexts teachers function in the dual roles of 'carriers' of standards, and as experts in making comparisons. If Condition C is achieved, some of the effort that is required to set up and maintain consensus for in-the-head standards (as under Condition B) can be rechanneled, allowing for more concentration on the actual processes of comparison.

On examination, it will be seen that Condition C is an indirect way to accomplish the same inter-rater reliability expressed in Condition B. Consequently, only one of the two is strictly necessary in an ideal situation. For sound, reproducible qualitative judgments, therefore, the requirements can be summarised as Conditions (A+B) or Conditions (A+C). (In a situation that is less than ideal, it is still fairly safe to place most of the emphasis on either Condition B or Condition C, provided that the other is not ignored completely.) Which one of Conditions B and C is preferable depends on a number of outside considerations. Initially, Condition C is almost certainly more difficult to realise than Condition B, but there are nevertheless some strong arguments for pursuing it as the preferred option. One reason is simply public accountability. If the standards in a subject are defined somewhere, they can if necessary be made available to persons not involved in teaching or the educations system generally. A second strong reason is pedagogical. If students know what the standards are, they know what they are aiming for.

Using a clearly defined system of criteria and standards as outlined under Condition C is unlikely to be more labour-intensive than conventional marking, once teachers become familiar with the process. Certainly one would not expect experienced teachers to be constantly referred to a book of standards for each appraisal they make. Once teachers get used to the idea of assessment by criteria, and accumulate experience in the practice of it, the standards will gradually become almost fully internalised. But the external standards specifications, rather than colleagues, would constitute the ultimate reference framework for assessment decisions. Inexperienced teachers would, of course, make intensive use of the standards specifications. It is likely that the use of explicit standards would considerably shorten the period of time it takes for beginning teachers to become competent assessors.

## Assessment policies in Queensland

How do these conditions relate to Queensland secondary schools? Condition A has been available ever since the abolition of external examinations. Under the Radford scheme, Condition B was achieved as district and state moderation committees worked towards achieving consensus on grades. At the back of the moderation procedures, however, was the technical requirement of achieving specified proportions of the seven grade levels over the state as a whole. Note that consensus can be achieved without the need to be explicit about criteria and standards. The ROSBA scheme, on the other hand, marked a move away from in-the-head standards (for these *did* exist despite the overt policy of norm-referencing, and *had* to exist for the scheme to work at all) towards explicit criteria and standards. In principle, the replacement of Radford by ROSBA corresponded to a shift from (A+B) to (A+C).

# Formative and Summative Assessment—A Complementary Approach

## Discussion Paper 6

**Abstract**: This paper attempts to describe where formative and summative assessment might most efficiently be applied under ROSBA. It touches also on one of the major concerns of ROSBA that summative assessment of students should not rely solely or even principally on one-shot examinations. In support of this concern, a rationale is presented for a series of student performances being used as the basis for judging whether summative assessments accurately reflect the real capabilities of students.

**Author**: Marilyn McMeniman, Assessment Unit, January 1986

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



## Introduction

Of prime importance in any discussion on assessment is the nature of the assessment itself, and the ROSBA report, not unexpectedly, shows great concern for how the evaluative process should operate. Consider, for example, the following extract from the report (emphasis added):

(**M13**) (Assessment for semesters 1, 2 and 3) …should become the responsibility of individual schools, being regarded as progressive assessments which are not certified by the Board but which should contribute towards the end-of-course or exit assessment …

(**M21**) In their assessment programs, schools should be encouraged to place more emphasis on **diagnosis with remedial feedback**, where results obtained do not contribute to end-of-semester achievement classification …

(**M23**) While choice of assessment techniques should remain largely a prerogative of the schools, the end-of-semester assessment in each subject should be based on information collected in a **variety of ways** (for example, assignment, oral work, tests, etc.). Overdependence upon regular testing or one-shot examinations should be discouraged and taken into consideration in the accreditation of school programs in each subject.

The above recommendations imply that teachers are expected to walk carefully between two sets of bipoles. The first bipole relates to the diversity of test instrumentation. Assignments, oral presentations, tests, etc. represent one end and the sole use of formal tests represents the other. Under ROSBA, teachers are advised not to favour unduly the more traditional assessment instruments. The second bipole is concerned with the purpose of testing. On the one hand, teachers are counselled to allow their Semester 1, 2 and 3 assessments to contribute towards the end-of-course or exit statement. On the other hand, they are required also to place more emphasis on diagnosis with remedial feedback where the results obtained do **not** contribute to end-of-semester achievement classification. Thus teachers under ROSBA are expected to (a) make use of a full range of assessment instruments, and (b) become involved in two distinct types of assessment: (i) formative assessment and (ii) summative assessment.

Formally, these terms are defined as follows:

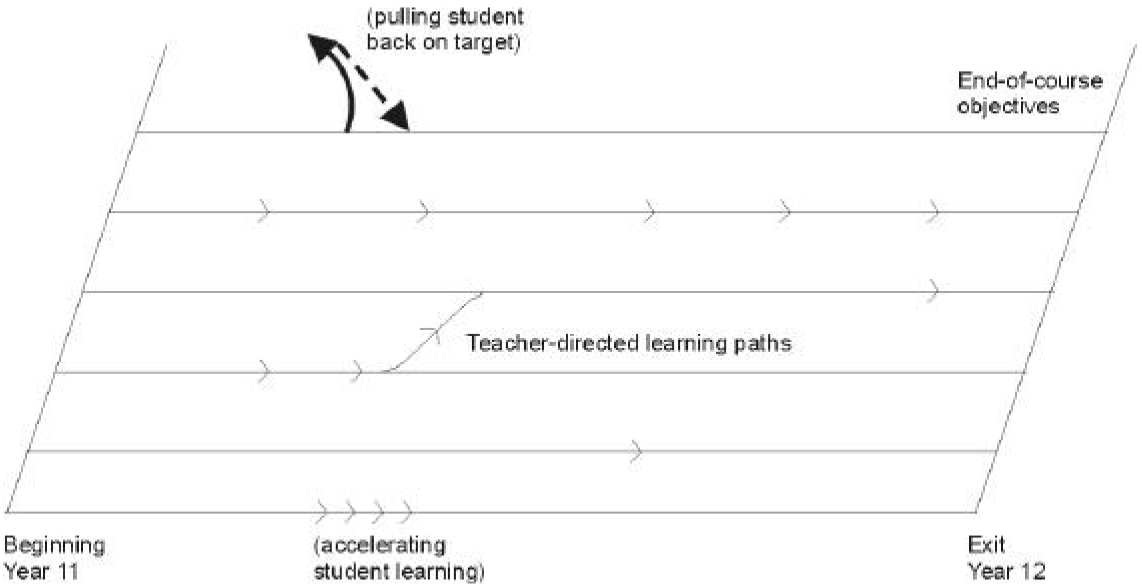
> **Formative assessment** occurs when assessment, whether formal (eg. testing) or informal (eg. classroom questioning), is (a) primarily intended for, and (b) instrumental in, helping a student attain a higher level of performance.
>
> **Summative assessment** is designed to indicate the achievement status or level of a student. It is geared towards reporting at the end of a course of study, especially for purposes of certification.

The question which teachers under ROSBA must address is which type of assessment is most appropriate and valid at particular points throughout the secondary years of schooling.

## Formative assessment

For **systematic improvement in student performance** of the kind envisaged by the authors of ROSBA, assessment that is designed specifically to be formative is the more effective. It is possible to think of a teacher's evaluative mandate as one of constantly pulling individual students 'back on target' towards the eventual realisation of long-term objectives. The following diagram relates to the role of the teacher as formative assessor. The pathways from Year 11 to Year 12 are plotted by the teacher and consist of a succession of increasingly sophisticated learning tasks that students need to undertake before 'arriving' at the long-term objectives.

### Figure 1: Examples of learning paths




The ROSBA work programs produced by schools show clearly that teachers are able to plan the learning experiences leading to achievement on different dimensions of particular subjects. For students to progress expeditiously towards their objectives, however, and to be pulled (or pull themselves) back on target, they need knowledge of their actual performance level, the reference performance level (that is, the level they are aiming to achieve), and an efficient way of pursuing their learning in the immediate future. It is in this last area that the teacher is indispensable.

There is good reason to believe that there will always be a formative role for the teacher that cannot be usurped even by computers despite advances in technology. Recent advances in computer assisted learning and adaptive testing make use of sophisticated strategies for returning information on performance to students. However, the teacher possesses knowledge not only of the student's current level of performance (the actual level) and the level of performance he can reasonably hope to attain (the reference level, whether reference is stated in short or long-term levels), but knowledge too of the individual attributes of the student which determine, in part, the nature of the learning experiences provided by the teacher. When assessing formatively, the teacher is concerned with the learning process rather than the product. That is, the teachers needs to be sensitive to how the underlying reasoning process of the student is working, rather than whether the answer the student produces is simply right or wrong.

ROSBA is a **school-based** system of assessment and, as such, the school is able to exercise autonomy in assessing and meeting the learning needs of its own student population. This autonomy allows for a variety of teacher responses to the differing learning patterns and rates of students. Although individual standards on different criteria may be objectives common to large groups of students, the paths different students take to get there may well differ. The person most capable of directing those paths is the teacher. Thus suggestions by the teacher as to how the gap between present achievement and the aimed-for standard can be lessened will, in all probability, differ from student to student.

When assessing formatively, the teacher is often involved in giving information on two reference points. The first is related to the specific assessment task and consists of information on what would have constituted a very good response to that task. The second reference point relates to the standards the student hopes eventually to reach after a course of study. That is, in commenting formatively on an essay, for example, a teacher may refer a student to examples of the best essays and comment on how the shortfall between the student's own efforts and those of the better students might be lessened. In addition, a teacher may provide individual students with information on what to do to put themselves more firmly on course towards their end-of-course objectives. In other words, students may get information on both the quality of their work in terms of the item itself, and also in terms of how their effort on that particular piece of work 'matches' with the broader long-term standards. In drama, for example, a student needs information not only on the quality of a single physical movement, but also on how that movement fits **successfully** into a larger dramatic segment.

## Summative assessment

As stated earlier, summative assessment is geared towards reporting at the end of a course of study and is used, in particular, for purposes of certification. Under ROSBA, summative assessment depends on 'the fullest and latest' information on a student's performance. Most Year 11 and 12 subjects are two year courses and the majority are developmental in the sense that expertise develops more or less continuously along a number of dimensions during the whole of the two year period. However, there are elements (particularly content-related ones) within certain subjects that may be 'completed' in Semesters 1, 2 or 3. For such elements, the fullest and latest information needs to be assessed summatively at an earlier date than Semester 4, and will, eventually, be 'counted' towards the exit Level of Achievement. For those elements within subjects which are developmental (particularly process and skill-related ones), summative assessment in the sense of fullest and latest is most appropriate and valid towards the end of the course. For students to 'piggy-bank' results from the beginning of a developmental course invalidates the assessments of two groups of students:

a) those for whom the penny drops late in the course, and who can never recover the 'losses' of the earlier semesters, although towards the end, they are achieving at the higher levels.

b) those who are very diligent workers and who accumulate good results in the early semesters but who find that the more difficult and sophisticated work in the later semesters results in an erosion of what were very promising achievement levels early on. As teachers under ROSBA are specifically requested not to allow assessment of the affective domain to contribute to an individual's exit statement of achievement, students' diligence should not be allowed to augment or offset a disappointing end-of-course achievement.

This is not to say that diligence is never assessed. Indeed, there is ample evidence in the research literature to show that time-on-task or attention-to-task is an important factor in achievement. Thus the effort expended by a student is reflected in the achievement level obtained.

**Overarching question**—As summative assessment is appropriate and valid only towards the end of a developmental course or after discrete units within a course, what does one understand by 'progressive assessments …which should contribute towards the end-of-course or exit assessment'?

The position taken in this paper is based on the work of Messick (1984) and researchers in the Australian field (Wood & Power, 1984) who have tried to deal with the discrepancy between what is actually assessed ('performance') and the underlying ability that gives rise to particular performances ('competence'). The theory goes something as follows. Suppose a student is asked to perform a specific task. What the student does, and therefore what the teacher assesses, is called a performance. But there are a number of reasons why the teacher's assessment of the performance may not reflect accurately the underlying competence—poor task specifications, distractions during the tackling of the task, unreliability in teacher marking, personal factors affecting the student (such as fatigue), and so on. The gap between the measure of the performance and the underlying level of competence can be decreased by making no testing or assessment episode crucial, that is, by obtaining a multiplicity of assessments of performance. In this way, the level of unreliability in assessment is reduced.

Under ROSBA, progressive assessments in a developmental subject can be likened to the teacher continually processing information on a whole range of student performances across a variety of tasks (oral presentations, assignments, formal tests etc.). This constant refining and updating of student performance enables the teacher to judge with a reasonable degree of accuracy what a student is capable of, and when particular performances are atypical. In everyday life, we do not judge people on one-offs or atypical performances. Rather we place our bets on more stable or well-established 'histories' that arise from consideration of several performances over a period of time. We know, for example, that one poor performance by an actor does not lead the public to conclude that the actor is incompetent generally, that is, beyond the incompetence manifested on a specific occasion in a specific performance.

Under ROSBA, teachers in Queensland have a unique opportunity to provide assessments that are informed by many student performances across a variety of tasks and over a considerable period of time. It should be possible, for example, to plot a student's development in a discipline by referring to examples of work over a number of semesters and thus give a stability or robustness to teacher judgments. On-going, formative assessment such as this can inform the end-of-course summative evaluation. (Incidentally, it can also help teachers make an **informed check** on the intervals between students for the SSA's.) Also for students who exit prematurely from a course, assessments that were intended for formative purposes will have to serve as the basis for an exit statement of achievement.

One final point—for formative assessment to work most efficiently, teachers need to take their students into their confidence and explain how formative and summative assessments are not mutually exclusive but **complementary** approaches to providing a **reliable** indication of student achievement. Teachers may well then get some respite from the question that students constantly ask—'Is this for assessment?'

## References

Messick, S. The psychology of educational measurement. *Journal of Educational Measurement*, 1984, 21 (3).

Wood, R. & Power, C. An enquiry into the competence-performance distinction as it relates to assessment in secondary and further education. 1984, Flinders University.

# Mathematics Criteria for Awarding Exit Levels of Achievement

## Discussion Paper 7

**Abstract**: This paper is about the criteria for judging student performance in mathematics at exit. The particular mathematics course considered is the current Senior Mathematics, although much of the paper has relevance for Mathematics in Society, and for Junior Mathematics. The first part of the paper considers some problems associated with current assessment practices, in terms of the syllabus and school translations of it. Then some contributions from different sources to the search for criteria are discussed. In the third section criteria are defined, and along with standards, organised as a possible model for awarding exit Levels of Achievement. The final section of the paper contains a discussion of the model and some possible implications. The model itself is included as a separate appendix.

**Author**: Janice Findlay, Assessment Unit, January 1986

***Note***: This discussion paper has been produced by a member of the Assessment Unit of the Board of Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



## Introduction

Under ROSBA, the contribution of each discipline to the secondary school curriculum is established in the Syllabus Rationale. The ROSBA scheme is a criteria-based one. In accordance with the above contribution therefore it would seem necessary that those criteria[1] are derived from the discipline itself, with at least some being *uniquely* characteristic of the discipline.

## The Present Situation

It is an inherent aspect of ROSBA that schools are required to translate the Syllabus into the school's particular context. One such element of the Syllabus is the statement concerning the award of exit Levels of Achievement (p. 34). This statement is reproduced below, and some features of it discussed.

### The Syllabus

The present draft of the Senior Syllabus provides definitions of the three categories of content, skill and process objectives (p. 2), and a statement pertaining to the criteria for the award of Levels of Achievement (p. 34) as follows:

[1] A criterion is a property, dimension or characteristic by which something is judged or appraised, as defined in the Discussion Paper entitled ROSBA's Family Connections by D.R. Sadler.

## Criteria for the Award of Levels of Achievement

### Figure 1: Criteria for the award of Levels of Achievement—Draft Senior Syllabus in Mathematics, 1983.

| | |
|---|---|
| **Very High Achievement** | The student should demonstrate a detailed knowledge of the content, a very high degree of accuracy in the application of definitions, formulae, and learned procedures, and a high degree of initiative and success in applying knowledge and techniques to unfamiliar situations. |
| **High Achievement** | The student should demonstrate a detailed knowledge of the content, a high degree of accuracy in the application of definitions, formulae, and learned procedures, and some initiative and success in applying knowledge and techniques to unfamiliar situations. |
| **Sound Achievement** | The student should demonstrate a general knowledge of the content, accuracy in the application of definitions, formulae, and learned procedures, and some initiative in applying knowledge and techniques to unfamiliar situations. |
| **Limited Achievement** | The student should demonstrate some ability to recall formulae and definitions, to apply formulae, and should recall and apply learned procedures in routine situations. |
| **Very Limited Achievement** | The student has gained some experiences in the course of study. |

The criteria referred to in the title are actually the five different *statements* themselves. However, within the statements (except for VLA) there are criteria of a different kind. These latter criteria relate more to the actual subject. Further, some references to standards are stated or implied. Rearrangement of the syllabus statements, to show separation of criteria and standards for each exit level produces the following table:

### Figure 2: Criteria and 'standards' implied in the Draft Senior Syllabus in Mathematics, 1983.

| CRITERIA | VHA | HA | SA | LA | VLA* |
|---|---|---|---|---|---|
| i. knowledge of content | detailed | detailed | general | some ability to recall formulae and definitions and learned procedures | – |
| ii. application of definitions, formulae and learned procedures | very high degree of accuracy | high degree of accuracy | accuracy | some ability to apply formulae and apply learned procedures in routine situations | – |
| iii. application of knowledge and techniques in unfamiliar situation | high degree of initiative and success | some initiative and success | some initiative | – | – |

*The category VLA has *no* criteria attached to it. To qualify as VLA, one 'has gained some experiences in the course of study.'

This rearrangement highlights a number of features:

(a) Criteria (ii) and (iii) differ only in the *context* in which the knowledge or procedures themselves are applied.

(b) The three criteria stated are not uniquely mathematical in nature. In their general form, they could characterise, for example, physics or chemistry as well.

(c) The terms used to denote different standards associated with different exit levels are also general in nature. ‘Detailed’ and ‘high degree’, for example, have not intrinsic meaning. They need to be defined. Enormously diverse interpretation by schools is possible with such general terms. For the purposes of statewide comparability, which is a goal under ROSBA, more precision is necessary.

(d) The category VLA does not involve any of the three criteria used for the other categories. In reality, it has no subject-related criteria attached to it. To qualify as a VLA, one ‘has gained some experiences in the course of study’. This statement gives no indication of what mathematical performances, if any, such a student is capable of. This is probably inevitable when there is not other lower grade available.

## School Interpretations

School Work Programs are required to provide assessment information (Items 8 to 12, Form R2, BSSS document). A *survey* of accredited phase II school Work Programs, (June, 1985) indicated that *almost* all of these documents contained the following:

(a) information on the frequency of testing and format of instrument;

(b) information on the weighting of different instruments and semester results; (c) descriptions of the styles of questions to be used (multiple choice and so on); (d) descriptions of content areas t to be used in testing;

(e) a restatement of the Syllabus Criteria for the Award of Levels of Achievement; (f) a representation of the criteria in (e) in a numeric form;

(g) a formula for combining semester results to obtain aggregates which will then be compared with the numeric cut-offs (at exit).

Very few schools indicated what peculiarly mathematical traits would be used to judge student performances. These traits are distinct from listings of content, which are used as vehicles for the development of such intellectual skills as logical thought and procedural knowledge.

A typical example of a numerical representation is shown in Figure 3 below:

| | Content/skill | Process |
|---|---|---|
| VHA | 85–100% | ≥ 75% |
| HA | 75–85% | ≥ 60% |
| SA | 55–75% | ≥ 20% |
| LA | 30–55% | – |
| VLA | 0–30% | – |

### Figure 3: A typical school’s numerical interpretation of the Syllabus Criteria for exit Levels of Achievement.

The numeric scores which replace ‘detailed’, ‘very high degree’ and so on, are intended to set minimum standards for each exit level. However, these numbers have not intrinsic meaning and depend, among other things, on the selection of items for the test, and the difficulty of those items. For the purposes of assessment, content and skill have been combined to form a singe criterion. This, together with process form the two criteria by which students’ scores are categorised. These two criteria are meant to match the criteria embedded in the Syllabus statement for award of Levels of Achievement. Neither descriptive nor numeric modes adequately reflect those mathematical dimensions that teachers purport to develop, and therefore, assess in students who study Senior Mathematics. Those characteristic traits of mathematics as a discipline which are to be developed at that level must be clearly identified if assessment is to be genuinely criteria-based. Of course, the global aims as indicated in the Syllabus pull together all these traits but not in a form that is directly assessable. The actual objectives are the 'learning experience' forms of these traits. While these objectives are used to judge students' performances at different times during the course of study, they are too specific to be utilised easily in determining exit Levels of Achievement after two years of study. The mathematical traits chosen to judge student performances must fit, *in general*, somewhere between the global aims and the course objectives. Further, these traits should exist as *commonalities* developed through several if not all of the eleven mathematics units. These conditions must be fulfilled if the exit Levels of Achievements are to represent composite pictures of performances in mathematics.

## Searching for Criteria in the Discipline of Mathematics

In an attempt to identify criteria that are characteristically 'mathematical', a variety of sources were examined. As well as texts from the wider field of mathematics itself and mathematics education (see bibliography), literature on educational objectives, particularly Bloom's Taxonomy of Objectives (see biblio.), the Syllabus (specifically the objectives), and general descriptions for each category were all analysed. The contributions of the above references to the search for subject-related criteria will be summarised.

Bloom has provided in his Taxonomy, a model for organising knowledge or objectives. Wilson[2] has adapted and utilised Bloom's Taxonomy of cognitive objectives, in his secondary school mathematics Table of Specifications which is used to select 'content' of assessment instruments. His adaptation consists of four hierarchically-related classes called: computation, comprehension, application and analysis, each class purporting to represent a more complex style of cognition.

The hierarchical nature of the four classes does not fit well with empirical evidence that many performance dimensions develop in students concurrently. Further the general nature of these 'criteria' does not allow them to characterise mathematics alone. An attempt by the author to write associated standards led to many very general terms, similar to the 'detailed' and 'high degree' in the present Syllabus statement for Award of Levels of Achievement. For these reasons, the mathematical categories of objectives, used by Wilson, appear to be unsatisfactory as criteria in mathematics.

The series of books on Lateral Thinking by De Bono focuses on the ability of people to restructure material so it may be perceived in a novel fashion. These books indicate that the mathematical style of thinking mainly developed at school level, typically focuses on a logical deductive (vertical) component, which is quite distinct from and in some cases antithetical to lateral thinking. (This is not to imply that lateral thinking is never used in mathematics, merely that the senior secondary school approach strongly develops the logical deductive component.) This fundamental deductive logic, being such an important and intrinsic aspect of the mathematical approach in senior courses, might be utilised as a criterion for judging student performance. Presumably its salience would be associated with VHA and HA performances.

From the literature search and analysis of the working involved in complex and/or difficult problems designed for Year 12 students, combined with the author's experiences, descriptions of the general characteristics of performances for each category were developed. Criteria and standards here were highly intertwined. Also obvious was that some criteria were salient for some categories but not others. These descriptions highlight a general 'technicianship' quality associated with a 'sound achievement' performance. This quality is defined as recall and reproduction of learned and practised content and procedures. VLA and LA performances evidence lesser amounts of this quality. The descriptions embody for HA and VHA performances an additional quality called 'mathematical ability' (for want of a better term). In showing this behaviour students utilise learned procedures in unfamiliar or novel situations.

[2] See bibliography.

Application, analysis, synthesis, and divergent thinking are all attributes of this generalised ability. Technicianship and mathematical ability are of course highly general terms, but seem to imply a number of criteria to be assessed: algorithmic repertoire, knowledge of language and notation, and some aspect of the logic referred to previously.

In the present Syllabus, the objectives are separated into content, skill and process categories *for each unit*. When each category of objectives is rearranged without reference to specific topics or units, the following themes within each category of objectives appear.

i. The content objectives can be divided into a number of sub-sets: *knowledge of:* definitions and meanings, techniques and procedures, formulae, theorems, and their proofs, properties, specific conditions or relationships, representations of concepts (e.g. equations and other), specific values of functions.

ii. The skill objectives also can be classified into a number of recurrent operational words: calculate, evaluate, substitute, simplify, find or solve, draw or sketch graphs, represent, apply or use, read or interpret.

iii. There are two major sub-sets of process objectives. The first group consists of developmental ones, covering all units, irrespective of content. These objectives focus on selection, application, analysis, interpretation, pattern searching, synthesis and evaluation (as used in its judgement context). The other sub-set covers objectives specific to the unit. However, these remain process areas by virtue of the 'complexity' despite having been taught in routine situations.

Restructured this way, the importance of such operational terms (or their synonyms) as: 'represent', 'apply these procedures', 'select', 'calculate', 'solve', and 'analyse' became apparent. These appear consistently throughout all units, and with significant frequency within units. It would appear therefore that they are mathematical dimensions on which students are required to achieve.

## Proposed Criteria for Mathematics

It is logical to consider that the criteria to be identified consist of the particular sub-set of all the discipline-related criteria which will be used to assess the performances of students at exit, after a two year course of study in mathematics at Senior level. This is in accordance with the relevant Queensland Board of Secondary School Studies Mathematics Syllabus. This set of criteria has to be sufficiently general to encompass all combinations of units of study, and reflect those general characteristics developed through the Senior Mathematics Course. Further, the exit Levels of Achievement must be able to be derived using these criteria.

Both the literature cited and general descriptions fall short of the needs of the Queensland ROSBA situation. Similarly, the Syllabus provides information which needs some support if it is to ease the operation of awarding exit Levels of Achievement, while achieving comparability. However, each of these, particularly the Syllabus, has provided valuable insights and assisted in the identification and definition of the following proposed discipline-related criteria for assessing student performances at exit. As a distillation of the criteria for assessment in mathematics, it is proposed that they be knowledge of language, representation, manipulation, procedure and attach mode. These criteria as defined below, are not hierarchical but are interrelated.

*Knowledge*: of symbols, signs, conventions, formulae, notation and so on, where the basic intellectual process is memorisation. The main testing technique involves stimulus-response (in reflex reaction fashion). Students may be called upon to respond by recognition, or by recall or supply. The main teaching strategy is rehearsal, either explicit or implicit through usage. (How much importance this type of testing would assume within assessment is another issue.)

*Representation* is defined as the dimension relating to translation: from one form of symbolism (language) to another. Specifically in mathematics, this would include verbal math symbolism such as graphs, diagrams and equations. Testing techniques and teaching strategies similar to those listed for Knowledge could be used. Students' responses include those as listed above for Knowledge, as well as more creative or constructive ones where testing situation involve generation of non-standard specific mathematical equations and graphs. (Where tasks involve routine procedures with information, that is not an exact reproduction of practised ones, but in the same class, this latter style of response will apply.)

*Manipulation*. This dimension includes both numeric and algebraic manipulations. It assumes a hierarchy between the numeric and algebraic components with algebraic being the higher. Further, the manipulation involves simplifications (easy and more difficult ones) which utilise factorisations, expansions, standard identities, and calculations. The operations involve essentially the four basic ones of +, x, -, , plus exponentiation. This criterion is associated with 'routine' or familiar situations, and their associated testing and teaching styles.

*Procedure*. This dimension refers to the use of algorithms which consist of finite routines of systematic or ordered steps, proceeding to an answer. It encompasses specific 'solving' as applied to equations with unknowns, but not the generalised 'solve this problem'. It also includes standard algorithms associated derivatives, integrals, equations and normals to lines, binomial expansions, solution of triangles and so on. Reproduction of the algorithm is dependent on recognition of familiar cues or commands. This aspect is used as a feature in identifying examples of the 'procedure' criterion. Increasing difficulty distinguishes the types of algorithms.

*Attack Mode*. This dimension relates to non-routine or unfamiliar tasks, whereas the previous criteria are *mainly* associated with routine ones. This criterion considers performances based on identification of clues in tasks, and transformation of these into specific operations. Important elements of this criterion are strategy selection and execution. Selection depends on identification of the strategy which best matches the cues and other information provided in the novel task, while strategy execution involves algorithmic procedures and associated manipulations. Interpretation, analysis and syntheses are also major processes involved in this behaviour.

No claim is made here that the dimensions above form an exhaustive listing of mathematical dimensions. Nor is there a claim that they encompass all dimensions on which student performances are judged within that part of mathematics taught at Senior level at Queensland schools. There are possibly other dimensions, for example, 'generalisation power' (involving pattern searching with actual number, and more algebraically described ones), as well as a dimension associated with spatial concepts, which are not included. However, as a first attempt, the Syllabus objectives appear to be encapsulated reasonably well by the proposed set. The global aims not directly associated with affective objectives are also reflected in these criteria.

## Development of an 'Exit Level of Achievement' Policy

Alone, the proposed criteria from the previous section do not provide sufficient information to derive exit achievement levels for students. Standards associated with these criteria must be employed for this task. The standards chosen belong to that section of the criterion continuum that considers *possible performances from lowest to highest* (ordered) at exit, after two years study of mathematics a senior secondary grades. Further, these standards are clear signposts
in students' performances. The 'overall standard' (Level of Achievement) is composed of allowable combinations of achievements on each criterion.

An attempt has been made here to select that chosen range for each criterion defined and also the identifiable, distinguishable pegs or standards within that section. An assumption has been made regarding the standards, namely that each standard is in effect a possible culmination point of a whole number of activities. Further, those learning activities may serve in some cases as interim standards, at prior points in the course. These interim standards (at end of topics, units, semesters and so on) and their relationship to the proposed exit standards are *not* explored here because decisions about them properly belong within the school.

The standards and criteria are presented using the model proposed by M. McMeniman in a Discussion Paper entitled *A Standards Schema*. Naturally the criteria, standards and the specific objectives for all the units must be consistent with one another. Some work matching the criteria and standards with these objectives from the 11 Senior Mathematics units has already proceeded: the purpose being to determine particularly the utility of the present standards chosen.

These standards are to be regarded as a first approximation only and it is intended that they be revised in the light of further experience. Despite this qualification, the concept should be clear.
The schema is presented as an appendix.

## Discussion of Criteria and Standards as Shown in the Model

The first four criteria are associated mainly with routine tasks, while the last one relates to non-routine tasks. All criteria consider student performances. However, the first four are expressed in terms of mathematical tasks. The last criterion describes the range of performances associated with attacking tasks. The order of the standards for the 'routine' criteria is assumed to represent mainly the increasing *difficulty* involved in the tasks, because it is assumed that routine implies no complexity. However, the higher standards in the

'routine' criteria still include some elements of novelty. The 'Attack Mode' criterion incorporates complexity. This difference between 'difficulty' and 'complexity' is important. The algorithm for finding stationary points of a curve is standard menu in Queensland Senior Mathematics classrooms. Nevertheless such a task can prove *difficult* when the associated algebraic expression requires mechanical agility and meticulous care to extract the required results from the expressions. Similarly, many of the minimum/maximum calculus tasks set as
'process' ones are merely more difficult examples of routine algorithms, which do not involve much 'hard' thought, often because the tasks set are reproductions of previously rehearsed ones.
This issue, the confusion between difficulty and complexity in 'process' questions, is basically another area of concern, but has been mentioned here because of the characteristics of the criteria and standards presented.

Complexity is related to the degree to which students can integrate ideas, analyse, abstract and create new relationships. It is inherent in the Syllabus statements about 'Powers of Analysis' and Process Objectives (p. 2). An attempt to measure indirectly this dimension has generated the last criterion.

The first group of 'routine' criteria could possibly be sufficient to distinguish VLA and LA performances from the upper levels. They could also differentiate between 'high' and 'low' achievement within a category. The first three standards on 'Attack Mode' could consolidate opinion about the three categories of VLA, LA and SA. The upper standards on this criterion could be the important ones which characterise HA and VHA, and confirm their differences from the SA performance.

The similarity between the structure of the standards for both 'manipulation' and 'procedure' is apparent. Manipulation, in the algorithmic sense, is so often a highly important aspect of procedure in mathematics. It may therefor seem that these are not sufficiently distinct to separate. For the moment they are left so, as there are some cases where manipulation is not seen as purely algorithmic. An immediate example is simplification of trigonometrical expressions using addition and factorisation formulae. However, the intermeshing of these two dimensions must be considered. The 'Attack Mode' criterion does not lend itself automatically to traditional, time-limit examinations. Nevertheless teachers' professional judgements derived from classroom observation and student-teacher interactions are often confirmed by these conventional forms of assessment. It should not be discounted merely on the basis that traditional examinations provide 'hard' evidence of student performance.

A most important question regarding selection of criteria and standards relates to whether they are practical. It could be argued that to some (if not a large) extent, these statements merely put into words and structure the thoughts of many teachers. The difference might only be that teachers think of these criteria, and particularly the standards, *in terms of actual problem and test papers*, and that this structure therefore initially appear highly novel. Further, is the range of performances selected truly indicative of those expected at exit? Whether or not this really is the case needs to be tested. Of crucial importance also are the criteria. Do they reflect Senior Mathematics, Vintage 1985, in Queensland schools? In asking this question, another is implicit. What should Senior Mathematics be? This has obviously been a source of much discussion for the Mathematics Subject Advisory Committee and will continue to be so.

The combinations acceptable for different exit levels is another issue. Obviously the particular combinations presented here are meant to be illustrative. They are not definitive. Rather than these minimum standards, could trade-offs such as 'S6 on one criterion with at least S5 on the others' make better sense in practice? Inevitably one will also question the role of 'marks',
and in particular SSAs for use in the derivation of TE Scores. This has not be addressed here, but is considered in a Discussion Paper by D R Sadler entitled *A Method of Deriving SSAs without Marks*.

## Implications and Future Directions

This paper is not meant as a final solution! It has presented some problems facing mathematics in regard to links among global assessment, the requirements of ROSBA, and the syllabus statements for those global assessments. Resolution of the actual *role* of Senior Mathematics has not even been tackled.
Further efforts by a working party of the Subject Advisory Committee might proceed in these areas:

(a) Cross-checking that the five criteria selected are a satisfactory coverage of what performance dimensions are essential in the mathematics course;

(b) clarification, refinement, and extension of standards within each criterion;

(c) referencing of these criteria and standards to all of the units within senior mathematics to see how universal is the application;

(d) a study of the utility of such a model for existing work programs;

(e) development of work programs using the model as a guide to the assessment section;

(f) effect on changing combinations of units to comprise, for example, Mathematics 1;

(g) usefulness in a flexible course of study (as defined in the Syllabus);

(h) development of interim standards for reporting to parents at required times within the two-year block;

(i) effect on summative assessment program from which exit levels derive;

(j) effect on assessment instruments;

(k) clarification of the differences between 'complexity' and 'difficulty' using examples from the different units studied.

## Conclusion

What is presented in this paper is an attempt at conceptualising mathematics and mathematics teaching in terms of traits or characteristics of mathematics *as a discipline*, including the subject matter, or topics and the usual manipulations that accompany this. This has been done to try to give substance to what criteria-based assessment should mean in mathematics under ROSBA. Many teachers who experienced the Radford moderation era would maintain that elements of criterion-referencing were practised at many of these meetings. That form however was oral and basically specific to the test instruments presented at the meeting. ROSBA requires a more holistic form which nevertheless still specifies sufficiently salient features of a student's performance in a discipline. Further, the ROSBA form must be formalised within Work Programs and the Syllabus, if comparability is to be pursued as efficiently as possible. This paper presents some preliminary steps in the process of identifying the discipline-related criteria for mathematics.

## References


Board of Secondary School Studies *Draft Senior Syllabus in Mathematics*, Reprint, July 1983.

Bloom, B. S., Hastings, J. T, and Madaus, G.J. (Eds.) *Handbook of Formative and Summative Evaluation of Student Learning*, McGraw-Hill: New York, 1971.

Costello, P., Ferguson, S., Slevin, K., Stephens, M., Fremboth, D. and Williams, D. (Eds.) *Facets of Ausralian Mathematics Education*, Australian Association of Matheamtics Teachers: Victoria, 1984.

Courant, R. Mathematics in the Modern World. *Scientific American*, September, 1964, 211, 3, 40-49.

Davis, P. J. and Hersh, R. *The Mathematics Experience*. Birkhauser, Boston, 1981.

De Bono, E. *Lateral Thinking: A Textbook of Creativity*. Penguin Books Ltd: Middlesex, English, 1970.

De Bono, E. *The Use of Lateral Thinking*: Penguin Books Ltd: Middlesex, England, 1971.

Henrici, P. *Elements of Numerical Analysis*. John Wiley and Sons Inc: New York, 1964.

Kline, M. *Mathematics in Western Culture.* Oxford University Press: London, 1953.

Kline, M. (Ed.) *Mathematics: An Introduction to its Spirit and Use*. W. H. Freeman and Company: San Fransisco, 1979.

Krulik, S. and Reys, R.E. (Eds.) *Problem Solving in School Mathematics.* National Council of Teachers of Mathematics Inc: Reston, Virginia, 1980.

Polya, G. *How To Solve It*. Doubleday and Company Inc: New York, 1957.

Quine, W. V. The Foundations of Mathematics. *Scientific American*. September, 1964. 211, 3, 112-127.

Sendov, B. L. Algorithms: The Algorithm stressed as a Central Theme in School Mathematics in Johnson, D. C. and Tinsley, J. D. (Eds.), *Informatics and Mathematics is Secondary Schools: Impacts and Relationships*. North Holland: Oxford, U.K., 1978, 53-55.

Trakhtenbrot, B.A. Algorithms and Automatic Computing Machines. D. C. Heath and Company: Massachusetts, 1963.

Wilson, J. W. Evaluation of Learning in Secondary School Mathematics in Bloom, B. S., Hastings, J. F. and Madaus, G. F. (Eds.) *Handbook on Formative and Summative Evaluation of Student Learning.* McGraw Hill: New York, 1971.

## Appendix: Criteria and Standards in Mathematics: Determination of Exit Level

| *Standards criteria* | ***S1*** | ***S2*** | ***S3*** | ***S4*** | ***S5*** | ***S6*** |
|---|---|---|---|---|---|---|
| Knowledge | When given explicit written or oral prompter or command recalls most definitions, theorems without proof and can state any given conditions; for majority of units studied. | With explicit written or oral command or prompter, recalls all definitions, theorems, given conditions, for all units; an can reproduce proofs of theorems. | Performs as for S2. Also, recalls symbolic notation, but does not always use this symbolism, when opportunity arises. | Performs as for S2. Uses whole repertoire of symbolic notation consistently, with ease and accuracy. | As for S4. In addition, accepts and uses new symbolic notation comfortably. | – |
| Represen-tation | Given standard graphical or geometrical representations, can verbally label these with associated mathematical equations, or describe characteristics. Can obtain obvious information from tables. | Given information, can translate routine (standard) mathematical equations into graphs or geometrical representations (inequalities, simple trig equations, parabolas, hyperbolas, exponents, logs, mechanics diagrams.) | (a) Given a specified equation, can manufacture other information to produce more difficult and/or complex figures or graphs (polynominals, rational functions, relative velocity, trig, conic sections).<br>(b) translate standard problems in written form into correct mathematical equations. | As for S3 and translate more difficult word problems or diagrams into mathematical equations or vice versa. | As for S4 and translate real-life problems (not text book ones as generally presented) into a suitable mathematical model (or approximation). | – |
| Manipulation | Perform direct calculation or simplification for numerical expressions or simple algebraic expressions | Perform basic algebraic manipulations (easy simplifications or factorisation) to obtain simpler form or answer. Amount of manipulation required is small. | Perform algebraic manipulations, where simplifications or factorisations are extensive. | Perform a number of progressive difficult simplifications (where progress cannot occur unless simplification is performed), to obtain end result | – | – |

| *Standards criteria* | ***S1*** | ***S2*** | ***S3*** | ***S4*** | ***S5*** | ***S6*** |
|---|---|---|---|---|---|---|
| Procedure | Perform algorithm with following characteristics; command is explicitly stated, one concept only, recall of formula with substitution followed by simplification, using 4 basic operations only. | Perform algorithm with these features: explicit command, formula recall, substitution followed by simplification: formula involves operations other than 4 basic ones, numerical/algebraic manipulation is only minimal. | Perform algorithm with: explicit command, one concept, however series of steps to be followed to obtain solution; algebraic manipulation necessary. | Perform ‘composite’ algorithm of this type: command explicit, solution obtained by application sequentially of a number of concepts. | Perform ‘composite’ algorithm with: explicit command, synthesis of a number of concepts previously developed, to obtain final answer. | – |
| Attack mode | Proceed with novel problem only when assisted through the crucial steps of the problem. Crucial links or patterns must be supplied. | Proceeds with problem, only when possible range of strategies are provided, and student trials these strategies. | Proceed with problems when clue in problem is indicated. Student then remembers the associated strategy/ relationship to use. | Independently extract clue from problem and translates this into strategy to be used. | Proceed with problem where no obvious clue exists. Is able to select strategy from problem information nevertheless. | Determine crucial steps in a problem and necessary information, when problem is open-ended and/or insufficient/too much information presented |
| | **S7** | | | | | |
| Attack mode | Identify a real life problem, its essential elements, any assumptions, and related mathematics | | | | | |

## Possible combinations table for awarding exit levels of achievement

| | *VLA* | *LA* | *SA* | *HA* | *VHA* |
|---|---|---|---|---|---|
| Knowledge | S1 | S2 | S2 or S3 | S3 or S4 | S4 or S5 |
| Representation | S1 | S2 | S3 or S4 | S4 | S4 or S5 |
| Manipulation | S1 | S1 or S2 | S2 | S3 | S4 |
| Procedure | S1 | S2 | S3 | S4 | S5 |
| Attack mode | – | S1 | S2 or S3 | S3 or S4 | at least S5 |

# Developing an Assessment Policy within a School

## Discussion Paper 8

**Abstract**: This discussion Paper contains a number of guidelines of potential interest to teachers and school administrators who wish to develop an internal school policy on assessment. The five principles outlined relate to the quantity and type of information required, the structure of subject in the curriculum, and the needs of students and parents for feedback. The Paper concludes with suggestions for two quite different ways of responding to a desire for reform.

**Author**: Royce Sadler, Assessment Unit, January 1986

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



The purposes of this Working Paper are (a) to raise awareness as to the degrees of freedom available under ROSBA in relation to assessment within schools, and (b) to outline a number of considerations that could be taken into account in devising a school assessment policy. The Paper does not deal with the desirable characteristics of assessment instruments or test items, the compilation of examinations, or marking schemes. There exist many books on measurement and evaluation in which these aspects are dealt with in detail. This Paper tackles some general policy issues, but is not meant to contain sufficient detail to serve as a step-by-step guide. It is clearly impossible in a general paper to deal adequately with the specialised requirements in different subjects. Some of the principles outline below are therefore more applicable to some subjects than to others.

In discussing assessment programs in schools, it is useful to make a number of distinctions. First, assessment programs in schools, it is useful to make a number of distinctions. First, assessments can be made under a variety of *conditions* which, for the sake of convenience, are grouped here under the two headings *formal* and *informal*. Formal assessment includes (among other things) all written tests, examinations, practical performances, projects and assignments, in which students are conscious that what they produce is to be assessed. Typically, all students are given the same (or equivalent) tasks, which are completed under clearly understood conditions. For example, students may not be allowed to communicate with one another. There may also be constraints as to the format, layout, length of the product, and time allowed. Informal assessment on the other hand is based on classroom questions, contributions in class, observations of student practices in the laboratory, homework completed, and what the student does (exercises, ‘seatwork’) during lessons. From these, the teacher builds up a composite picture of what the student is achieving through data gathered informally.

The second distinction is between *formative* and *summative* assessment, and relates to the *purposes* rather than the conditions of assessment (see McMeniman, 1985). Formative assessment is primarily intended for, and instrumental in, helping a student towards a higher level of achievement than that attained previously. It is specifically geared towards improvement, and the information is used both by the teacher (clarifying and revising work), and by the student (deciding to study weaker areas). By contrast, summative assessment is designed to indicate the achievement status or performance level of the student at the end of a course. Although data may be gathered throughout a course, summative assessment is geared not towards diagnosis and remediation, but towards reporting, especially for purposes of certification.

Under ROSBA, the possibility exists for data from informal assessment to contribute towards the determination of exit Levels of Achievement, although this option is not often taken up. There are several reasons for this. First, general impressions of what students are doing, particularly with respect to their sincerity and diligence, are not always matched by the quality of their achievements. Teachers can be misled, especially in the early stages of a course, by attentiveness, the ready response, and the appearance of industry into believing that achievement automatically must follow. Formal assessment plays a useful role in digging beneath the surface. Second, informal assessments typically do not give rise to 'hard' data, which can be produced as evidence in discussions about student achievements. Third, informal assessment (by definition) relies on information gathered incidentally rather than systematically. In situations where the quality of what students do is directly dependent on the circumstances under which the students attempt the tasks, informal assessment may be strongly influenced by chance.

Perhaps it is not surprising then that informal assessment is often associated with formative purposes (where the three limitations just mentioned are not so crucial), and that formal assessment is mostly associated with summative purposes. Under ROSBA, these associations could profitably be relaxed to some extent. With care, some assessments that are made informally should find their way into summative decisions. In a system of criteria-based assessment, Levels of Achievement should be determined by comparing students' achievements with defined criteria and standards, not with those of other students, as under Radford. Informal assessment can provide corroborative data to help in the classification. Furthermore, there is no necessity to use all formal assessment for summative purposes.

A key issue in the design of school-based assessment programs is the *amount* of testing or examination required. On the one hand, teachers often complain about the drudgery imposed by assessing students frequently, the energy required to set and mark student work, the time that assessment subtracts from the business of teaching itself, and the negative effect it has on teacher-student relations. Assessment seems to drive much of what goes on in the classroom. Fifteen years ago it was external examination, now it is school-based assessment. The philosophies and practices of the two systems are of course different, but it appears that assessment is still largely in control.

On the other hand, many teachers somehow feel obligated to test continuously, and to accumulate the results. Among the explanations proffered by teachers for behaviour of this type are (a) the students will not take seriously anything that does not count towards a final result, (b) it is necessary to be seen to be diligent and conscientious in assessment, as a form of professional insurance against complaint, and (c) the Board requires it.

It is worth mentioning in passing another factor that perhaps has contributed in a minor way. End of term examinations have been a persistent tradition in secondary education. That was the case under external examinations, it remained so under Radford, and continues to be so under ROSBA. Since the time several years ago when a decision was made to alter school holiday patterns, there developed a feeling that it was somehow 'right' to schedule tests or examinations before *each* vacation period. Two things followed. First, mid-semester tests in some schools assumed greater significance, taking on many of the trappings of semester examinations. Second, the mid-semester tests were held a little before vacation, so that teachers would have time to mark student scripts and get the results back to students before school went into recess. As a consequence, valuable teaching weeks were lost.

It is appropriate at this point to examine the Board's official attitude and policy on assessment, in order to examine the Board's official attitude and policy on assessment, in order to dispel some misunderstandings about the Board's position on continuous or progressive assessment. The Radford and ROSBA systems have many points in common, the most prominent being that both are school-based in their assessment of students. Historically, the move from external examinations to school-based assessment was due to a reaction against the pressure-cooker nature of one-shot examinations and against a narrow syllabus. The alternative to one-shot assessment is, of course, to have assessment spread over a considerable time period and a number of assessment instruments. That is what ROSBA is about. Progressive assessment does not mean that everything has to count, or that assessment should be incessant. Many schools are testing for too frequently.

It might come as a surprise to some teachers to find that one of the criticisms of the Radford system was the inroads made on teaching be assessment programs. It was the intention of the authors of the ROSBA Report that the new system would lead to a reduction in the emphasis on testing as such, but lead to a greater focus on teaching for improvement. Many schools seem to be under the impression that they must show in their Work Programs how performance on *each specific objective* is to be assessed. This moves them, in the interests of safety, towards the maximum number of tests tolerable in the school. But in a limited time, it is impossible to test all behaviours adequately. Sampling is inevitable. The question to be answered is this: How large a sample is necessary for a school to be able to decide, with an acceptable degree of certainty, on exit Levels of Achievement, bearing in mind the needs to assess a representative sample of the objectives? The amount of testing to be carried out in a school is primarily under the control of the school and its teachers and the Board is generally supportive of a reduction in the amount of assessment.

Assessment practices have a great influence on what students put their time into. The greater the demands made by one subject, the less time students have available to put into another subject. In some schools, a combination of heavy assessment requirements and a forceful (if unwittingly selfish) teacher can control the time allocations of the students. This issue is one that can be taken into account in a school policy on assessment.

A program of frequent formal tests, even if some are conducted wholly within the space of a class period, is a very effective tool for chopping up the school year. The result is that many weeks of the school year are lost to teaching through tests and examinations. A useful first step in developing a school assessment policy is to carry out an internal audit of the total investment of time and resources (by both students and teachers) in testing. This serves as a firm basis for working out a school policy, so that there can be co-ordination across subjects, and no overload in any particular subject. The policy itself could include specific guidelines as to: (a) how the school is to implement the Board's requirements for exit level assessment, taking into account that exit assessment should reflect selective updating, be based on fullest and latest information, and achieve a balance over the course of study, (b) the absolute amount of testing that is advisable within the school, (c) appropriate upper limits for each subject, and (d) a due-dates schedule to ensure that students do not have a number of assignments and projects all to be submitted for assessment during the same week. In the next section of this Paper, five principles for use in formulating a school assessment policy are enunciated.

## Principle 1: Information Need

Taking a cue from information theory, a useful distinction can be drawn between *data* and *information*. Data are the raw facts of a matter. Information on the other hand is connected with what use is or can be made of data. In particular, information can be thought of as anything that reduces ambiguity or uncertainty about a situation.

In theory, one might imagine that the more data one has the better. In practice, however, the law of diminishing returns applies. Data beyond a given point carry little information, that is, they cease to inform and are technically redundant. At the other extreme, small quantities of data may carry little information. The 'signal' of potentially useful information is contaminated by 'noise' due to measurement error and inadequate sampling. Figure 1 shows the general relationship between information level and quantity of data. Beyond the point marked *X* on the curve, no amount of additional data has much informational value.

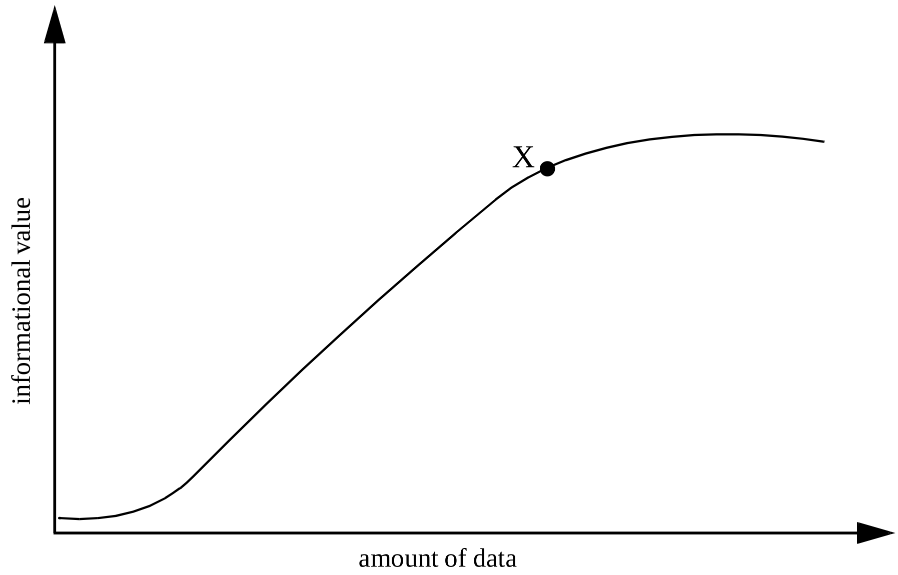



**Figure 1. The relation between quantity of data and its informational value.**

Now think about this in relation to the assessment of students. Incidental snippets of data, such as a teacher picks up in informal exchanges in the classroom, may not produce enough valid and reliable information on which to base a conclusion, although in some subjects, creating an appropriate response to a real situation that is previously unrehearsed is precisely what has to be learned. (Foreign language acquisition provides such an example). But by and large, students have to be set a variety of tasks so that their achievements can be displayed and recorded. Under ROSBA, it is intended that assessments be made by using a range of test instruments, formats, and situations. All of these are potential sources of data, and the aim of the teacher should be to generate only that quantity of information as corresponds to approaching point *X* on the graph from the lower side. Any more than that constitutes overtesting.

The first principle can now be stated explicitly: the amount of assessment should be just sufficient to reduce ambiguity about the achievements of students to an acceptably low level. This does not, of course, imply that non-assessable tasks should not be set for students. Most learning occurs through doing, and through getting feedback about the weaknesses of what is done and how it might be done better. That is the very essence of formative assessment. The first principle does imply, however, that for summative purposes, the amount of assessment should be limited strictly to what is necessary for a clear and reasonably complete picture to be obtained.

## Principle 2: Triangulation

If the first principle comes from information theory, the second comes from surveying and navigation. It is the principle of triangulation. To get a fix on a position on earth, several different readings are taken in such a way that they converge on the point in question from different directions. Good assessment practice is to aim for relatively independent readings as to exactly what a student can and cannot do. Student achievements can only rarely be assessed thoroughly by using a single form of test instrument.

External examinations were largely confined to pen and paper testing (with some exceptions in foreign languages and one or two other subjects), and laboratory practical books in the sciences (for checking only, not for grading). It was impossible to include data derived from projects, assignments, field trips, and library research. Under Radford and ROSBA, all of these can serve as data sources, which means that the scope of what is assessed is broad, and therefore that attainments not easily assessable by external examinations received due recognition.

The greater latitude under ROSBA is not based on a philosophy of variety for the sake of variety. It is a matter of choosing the most appropriate instrument for the purpose in hand. Certain types of achievements are best and most fairly assessed under controlled conditions, others are not. There is no logical reason why a school assessment policy should not acknowledge that formal tests may be appropriate in some subjects at particular stages during a course but not in others. Blanket expectations or requirements are seldom likely to suit all subjects equally.

## Principle 3: Subject Structure

The third principle is that the assessment program should reflect the structure of the subject, and the way it is taught. In some subjects, students begin with almost no experience and develop expertise gradually, over the whole two years of a course, on a number of fronts sat once. For the sake of convenience, call these *developmental* subjects. Again, foreign languages provide an example. Students typically do not learn to listen in one semester, speak in another, read in the next, and cap it off with writing. They pursue these parallel fronts simultaneously, not least because the fronts inevitably interact with one another. In other subjects, the material comprising the course is divided into semester units, which are intended to be nonsequential, and theoretically could be taught in any order. *Unitized* would be an appropriate label for subjects having this structure.
It should be obvious that for purposes of final reporting of achievement in developmental subjects, the critical data should be collected toward the end of the course. All the early data should be regarded as primarily formative, not summative. Other things being equal, it may be appropriate for there to be no formal tests at all for the first semester in a developmental subject! Data obtained informally during classroom interactions, from homework, or through class quizzes may be quite sufficient. Teachers should not shrink from choosing this as a realistic option. However it should be a deliberate, conscious decision, not one made by default. In such subjects, most of the relevant data for arriving at an exit Level of Achievement should be obtained during the final semester. Unitized subjects present a quite different case. Results have to be accumulated systematically during the whole course, and this normally implies that data have to be gathered at or towards the end of each unit. Of course, many subjects are neither wholly developmental nor wholly unitized. The fullest and latest information will then be based on some data gathered throughout the course, and other data gathered at the end. Depending on the structure of a subject, some of the formal assessment (or even all of it in they early stages of some courses) could well be reserved for purely formative purposes.

## Principle 4: Pedagogical Needs

The fourth principle is related to pedagogical needs. Learners need to have feedback on how they are doing, if only to patch up weak spots. They need progress reports. *Regular* testing is sound pedagogically when students are making steady progress. But in some subjects, especially developmental ones, there may be periods during which most students make relatively little progress either because the students reach some kind of plateau, or because this is characteristic of the subject at that stage. Not until this material or skill is consolidated can further progress be made, and when it does occur, such progress may take place rapidly. Assessment during the consolidation phase may be almost totally informal. As a general principle for longitudinal growth, the slower the rate of development the less frequent the testing need be. There is simply not much additional information to be gained by testing frequently. Conversely, the frequency of testing needs to be greater when change accelerates.

### Principle 5: Reporting

The fifth and final principle outlined here has to do with the necessity for reporting to parents, who quite naturally want to know how their children are progressing. It is administratively convenient to have all students in a school take their tests at the one time and to issue mid-semester and semester reports. But some creative experimenting in reporting to parents is probably called for. Parents do, of course, want to know early in the piece whether their sons and daughters are performing as well as can be expected. It is not at all clear why formal tests have to be conducted in every subject every eight weeks, even if school reports are a requirement. Perhaps it would be to the long-term benefit of both teachers and students if the coupling of school reports with the timing of tests and examinations were not so tight.

Parents often demand concrete data (such as test scores and place in class). They are not content with vague statements about satisfactory progress. This poses no insuperable problems *except* in subjects where, for a variety of reasons, test scores are quite inappropriate at a particular stage in the course. In that case, parents deserve to have explained to them when, how, and precisely why nonstandard reporting will be made. Such information could well be incorporated into the school prospectus. Interstudent comparisons (rankings) can be avoided, for example, by indicating to parents precisely where, on the performance continuum from virtually no expertise to almost perfect performance the student is currently placed, and the informal and formal evidence for the placement.

## Two Radical Proposals

In contrast with the earlier parts of this Paper, this section outlines two radical proposals for changing the approach to assessment in a school. Although somewhat speculative, they are intended as serious options, not as frivolities. To the criticism that neither would stand a chance of getting past a Review Panel, a few comments need to be made. Review Panels are not composed of people outside the education system. The vast majority of Panel members are practising teachers. Although they work to implicit criteria and rules in the accreditation of Work Programs, members of Panels are, like most teachers, receptive to new thinking on pedagogy. Again like most teachers, they are likely to resist novel ideas when such are presented without sufficient evidence of serious and sustained thinking. Any school proposing a nonstandard solution to a problem has two tasks before it, not one. The first is to present the solution, and to make certain that it is a true solution and not just wishful thinking. The second task is to convince the Panel that sound educational arguments are being advanced for the proposal. The Panel’s arguments against the proposal have to be anticipated, and answered in the accompanying documentation. Additional weight is given to a proposal if it can be shown that a pilot study of the assessment policy shows promising results. In convincing a Review Panel of the educational desirability of an assessment program that departs from tradition, the second task, namely the selling of the system, is at least as important as the first (and historically the one most likely to be either ignored, or carried out perfunctorily, by would-be innovators).

### Proposal 1

Cut the absolute amount of formal assessment by 50% across the board, in all subjects, apportioning the remaining 50% according to the five principles stated earlier.

Apart from dealings with Review Panels, there are a number of interesting questions that spring to mind. Would students from a school which adopted this proposal be disadvantaged in teachers’ meeting and consortia? Why would that be so? Would students work as hard if assessment were not so important (and continuous?) What other incentives could be constructed? Which types of incentives are educationally the more desirable? Could teachers get sufficient information from informal assessments to guide individual and group learning? Would competition increase or decrease? Would the extra weeks released for teaching more than compensate for the loss of systematic information about student achievement? Would more continuity in the teaching year lead to more sustained effort, and greater achievement in the longer term?

### Proposal 2

Keep the absolute amount of assessment the same, but improve the quality of it so that it becomes more facilitative of learning. The aim would be to turn assessment more directly into a teaching medium, replacing to some extent expository teaching to whole-class audiences.

The intention in this proposal is to turn assessment on its head, exploiting its peculiar power to drive and organise what students do, (including how they utilise resources, their time, and the teachers) and in a sense to help cater for individual differences. As a scheme, this proposal would be more appropriate in Years 11 and 12 than in the junior school. It would require that teachers define goals and tasks clearly, specify in considerable detail the criteria and standards to be used in assessing the quality of student work. and then encourage students to

(a) engage in evaluation (of, for example, work of their peers) as a vehicle for their own development, and

(b) work cooperatively, and also independently, in achieving the goals set.

### Transition

Either of these proposals would represent a major break from present practice in most schools. Neither should be attempted without fuller consideration of the risks involved, and whether these risks out-weigh the potential pay-offs. But either proposal could, if implemented, result in better education for students. The areas of greatest impact on students and teachers need to be identified, and contingency plans adopted. For example, in many schools, students refuse to take seriously anything that does not count. Students continually ask “Is this for assessment?” It is quite probable that students want everything to count for summative assessment not from absolute necessity, but because the education system has conditioned them over a period of a decade of schooling to think that way.

And that conditioning, being quite thorough, would take some undoing. In the present context, students would have to be re-educated, gradually, into seeing assessments not primarily as a means of banking credits and eventually withdrawing them for an exit Level of Achievement but as a means of knowing themselves, of mapping their own progress towards well-defined goals, of engaging in realistic self-assessment, and ultimately as a means towards the production of performances which earn them the appropriate Level of Achievement. It will obviously be necessary to back away from the present ethos of accumulating credits, and focus more on the overall quality of achievements. Such transitions do not come easily, or quickly, even for teachers. It has taken years for the present system to consolidate. It would similarly have to be dismantled slowly and carefully.

## Conclusion

It is common for schools to overtest, even though it is within any school’s power to reduce the amount of testing, and change the style of it. The five principles outlined here are intended to assist administrators and teachers who wish to develop an assessment policy within their schools in such a way that teachers' loads are reduced, more educative intentions are achieved, and relations between teachers and learners are improved.

# General Principles for Organising Criteria

## Discussion Paper 9

**Abstract**: Criteria are, by definition, fundamental to a criteria-based assessment. This Paper outlines a conceptual framework for organising criteria, using a particular subject as an example. It also shows how student achievement can be recorded using criteria.

**Author**: Royce Sadler, Assessment Unit, April 1986

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



In a criteria-based assessment system, criteria are, by definition, key elements. Once criteria are identified, however, it becomes essential to have (a) some conceptual framework for organising them theoretically, and (b) some practical scheme for recording student achievement in terms of them. The first part of this Discussion Paper explains how criteria may be elaborated, arranged, and displayed at several levels in what is essentially a hierarchy of terms. The second part shows how criteria from one particular level of a hierarchy can be organised to provide useful and convenient records of student achievement.

The discussion takes Senior English as a concrete example, but the guiding principles are quite general and can be applied to other subjects and to other levels of schooling. English has been chosen to illustrate the principles because the language of instruction in schools is English and thus every teacher is, to some extent, a teacher of English. But the particular treatment given to the subject here has not been endorsed by the English Subject Advisory Committee, and therefore has no official status. Furthermore, the concern here is less with the particular tasks which students in English may be asked to attempt than with the *criteria* that are used in judging the quality of students' work.

## An elaborative hierarchy

The basic proposition is that in general any given criterion can be either expressed as a component of some higher level criterion (moving up the hierarchy) or a broken down into a number of lower-level criteria. The reader should at this point refer to the chart on page 9, in particular to the criteria shown at Level 2. The four criteria listed all apply to *written* English, and are probably sufficient to cover most writing tasks. Organisation, presentation, and so on appear in many of the lists of criteria suggested in the literature for use by English teachers. The labels may, of course, vary but the intent is essentially the same.

Consider the criterion *Language*. Language is, according to the chart, and umbrella term covering spelling, punctuation, paragraphing, and several other properties shown at Level 3. Some of the Level 3 criteria, such as originality and register, could be further decomposed, if necessary, to form a Level 4. The criteria shown at Level 3 on the chart may not constitute an exhaustive set. Others may have to be added, either generally, or to cover assessment in specialised tasks. The Level 3 criteria provide meanings and interpretations for each of the criteria at Level 2, which in turn give substance or content to the criteria at Level 1. In practice, lower-level criteria may not be altogether independent of one another because the meanings of words have somewhat fuzzy boundaries, although an attempt has been made to avoid overlap as much as possible in the chart. Observe that developing a hierarchy downwards by specifying meanings for a primary criterion is a matter of interpretation and elaboration. Each lower level amplifies the level immediately above it.

The inclusion of Level 0 (Level Zero) on the chart serves as a reminder that *student achievement*, and not diligence, persistence, or attendance, is accepted as the sole criterion for awarding exit Levels of Achievement. Syllabus documents typically include several general assessment guidelines or conditions for the fundamental or Level) criterion. The English syllabus, for example, indicates that assess-ment should take place in a range of contexts, with a range of purposes, and for a range of audiences. It should, in addition, lean towards realistic rather than decontextualised tasks.

Just as a criterion is given content by expressing it in terms of a number of simpler criteria, its connection with a higher-level criterion is brought to light by asking why a particular criterion is thought to be worthwhile, that is, by calling for a rationale. For example, if one asks why register, which includes the choice of appropriate language, is important, the reply could well be that proper use of register establishes links with an audience, and communicates in a way that may be both more potent, and more specialised, than another set of words with the same semantic content. In other words, register is an important criterion because it signifies that the writer knows how to match the characteristics of the communication with some of the characteristics of the audience.

What follows are some points and suggestions which make a number of implications for syllabus writers and practising teachers a little clearer. Bear in mind, however, that the treatment of English as a concrete example here is meant to be illustrative, not definitive. It will be necessary to overlook any errors and inadequacies in order to get the overall thrust of the argument.

The closer the assessment system operates to the Level 0 criterion, the more the criteria are inaccessible to students, and the more teachers operate using intuitive or unarticulated criteria in what might be called ‘connoisseur-mode’. In this mode, although teachers make evaluative judgments (and they can be trained to make valid and reliable judgments), only the *results* of these judgments are available to students (except in so far as teachers’ written comments on student work refer to or imply explicit criteria). Criteria-based assessment operates most effectively further down the hierarchy, so that the criteria in use at any one time, or for any one assessment episode, are visible to teacher and student alike. However, there are dangers in getting too atomistic. The more detailed the criterion lists, the more unwieldy they become, and the more difficult the job of assessing students. Taken to the limit, a multitude of little criteria inevitably starts to look like a checklist, the slavish use of which substitutes separate fragmentary judgments for fuller, more integrated judgments.

Syllabuses should include both Level 2 and Level 3 criteria, and should also indicate the relation between them. The Level 3 criteria (which could conveniently be called ‘sub-criteria’) are, or should be when they are properly worked out, implied by those at Level 2. Level 3 criteria are important to both teachers and students, partly because of their obvious role in defining the Level 2 criteria, and partly because of their utility in formative assessment. For example, students may find it easier to appreciate the significance of teachers’ comments about handwriting, visual appeal, and layout than to see the relevance of more general comments about presentation, which is a Level 2 criterion. Although Level 3 criteria are often, however, too fine-grained to feature strongly in the formal criteria-standards specification for the award of Levels of Achievement, they should nevertheless be mentioned specifically whenever they constitute important, non-negotiable elements.

Some of these Level 2 and Level 3 criteria may appear in syllabuses other than English, and so they should. But the point is that a collection such as the one in the chart should capture the essence of English (in Queensland secondary schools, in the 1980’s), and distinguish English from other subjects.

In constructing the Level 2 criteria, an attempt has been made to keep to a manageable number (which is probably between four and ten), and have them all of the same ‘order’ so that fine-grained and coarse-grained criteria are not mixed together at the one level. On the latter point, suppose it is considered that, say, *organisation* and *legibility* (two of the criteria for writing contained in the present syllabus) are somehow not commensurate, because legibility, although important, seems to be at a lower level than organisation. To reflect this thinking, legibility (or handwriting) has been made a component of *presentation* in the chart.

Over 40 criteria for written English have been identified in the literature on the assessment of writing. Among these there is, as one would expect, considerable redundancy. Apart from the fact that they exist at various levels in the hierarchy (so that some subsume others), use of such a large collection is for practical reasons out of the question. Furthermore, words when isolated from a context have a natural imprecision. All of these factors point to the conclusion that a definitive list of criteria for assessing achievement in written English is probably a vain hope. What should be achievable, however, is consensus on a working framework.

There is, furthermore, no unique decomposition of the major criteria. It is possible to construct equivalent sets of criteria, equivalent in the sense that two or more sets may ‘cover’ the same ideas, even though in different ways. This does not imply, however, that different but equivalent sets are equally useful. A criteria-based assessment system in which teachers use sets of criteria that are more or less peculiar to their schools makes it difficult for teachers from different schools to communicate with one another. It is important, in the interests of dialogue, to encourage the use of a common vocabulary, incorporating as far as possible terms that are already familiar to teachers of the subject. The use of a consistent terminology is also important for students who seek to improve the quality of their work. Without it, continuity in using criteria over the period of a two year course (particularly if there are several teachers) may be hard to achieve, the meanings and practical implications of at least some criteria remaining mysterious.

If consensus is accepted as a worthwhile goal, however, there are two further aspects to be kept in mind. The first is that the criterion set should be chosen carefully, so that it does in fact cover in one way or another most of the salient dimensions. For example, one criterion that features in some lists of criteria for assessing written English is *flair*. It does not appear in the chart. The question is whether what is already there is sufficient for addressing that criterion, given that for some types of writing it may not be applicable. It could be argued that flair is, for certain classes of student work, implied by the following collection of Level 3 criteria: originality, coherence and register. If that argument is sound, there would be no need to list flair as a separate criterion. (Two other criteria to think about are aplomb, and flavour.) The second aspect is that while the criterion set should aim to be useful for the majority of types of writing, and be applicable to the majority of student submissions, there may occasionally be a need to invoke criteria not specified or implied in the chart in assessing pieces which are exceptional in some unexpected way. That should be seen as normal, and not an aberration.

The STANDARDS for the various Levels of Achievement in English could probably be satisfactorily conveyed by something like the following:

(a) General statements of criteria and standards, preferably though not necessarily set out as in the cross-tabulated standards scheme described in the Discussion Paper, *A Standards Schema* (McMeniman, 1985).

(b) Several exemplar folios for each Level of Achievement, showing a variety of legitimate interpretations of the syllabus.

(c) Annotations of each piece in each folio (several sentences) giving details of the *conditions* under which each was produced (for example, examination, or take-home project), and drawing the reader’s attention to features of the piece that distinguish it as good, bad, or indifferent, and justifying its inclusion in the folio.

(d) Annotation of each folio as a whole, giving reasons for its nominated classification, and drawing attention to such features as the consistency or inconsistency (peaks and troughs) of the quality of work in the folio, and the trade-off decisions (if any) that were made.

The annotations about the quality of student work should, as a general rule, emphasise criteria from Level 2, although reference to some Level 3 criteria may well be necessary when distinguishing various points (standards) on a single criterion. If, however, there are some non-negotiable elements from Level 3, these should be mentioned specifically. Spelling, for example, might be an aspect singled out if the syllabus writers have a policy on the matter. There are likely to be some other special cases where it would be both natural and beneficial to refer to some Level 3 criteria. The main point is that teachers should be encouraged to think about and understand thoroughly what assessment in English relative to published criteria and standards means, and what it implies for practice.

## Recording achievements using criteria

At this point it is appropriate to investigate a recording system in the school so that performance on the various criteria can be tracked. The comments here refer primarily to summative assessment. For each type of writing, all of the Level 2 criteria are considered appropriate. A simple and convenient method of recording is to open and maintain an achievement table for each student. For example, suppose that a school, in a certain period of time, requires its students in English to complete a cartoon, a poster, a character portrait, an argumentative essay, and a dramatic script. The following table shows how performance of one student might be recorded:

| | **Content** | **Organisation** | **Language** | **Presentation** |
|---|---|---|---|---|
| cartoon | X | X | X | X |
| poster | X | X | X | X |
| character portrait | X | X | X | X |
| argumentative essay | X | X | X | X |
| dramatic script | X | X | X | X |

Each X in the table would by the end of the period contain a record (preferably non-numerical) of the standard of the student's work on the associated Level 2 criterion. With practice, it should become as simple to record using the table as it is to assign marks and record them in a mark book.

Although the table shows each of the Level 2 criteria being used for each task set in written English, this is not meant to imply that all subjects will follow the same pattern. In some subjects, some boxes would remain empty for the simple reason that not all assessment instruments can be used to assess on all criteria. Deciding which boxes are to be 'live' for a particular assessment program is part of the planning of that program.

Even in English, not all of the Level 3 criteria are equally important for all tasks. For example, originality is likely to be important in all cartoons, but may be less important for some types of argumentative essay. Paragraphing is simply irrelevant to posters and cartoons, and illustrations may not feature in dramatic scripts. The principle is to 'pick up' the relevant Level 3 criteria appropriate to the various tasks, choosing the tasks so that overall most (if not all) of the Level 3 criteria receive attention somewhere in the assessment program. Note too that originality may be highly important for a particular argumentative essay, but not for another. The form or genre of the task (poster, cartoon, essay,...) is not the only consideration; the requirements of a *specific* task have to be taken into account. That is, there may not exist a set of particular criteria that should always be used as 'task criteria' in assessing argumentative essays, but a particular argumentative essay set for a group of students may provide the teacher with an opportunity to assess on a particular set of dimensions. In practice, of course, the students would be told (possibly, in English, by means of what English teachers call a 'criterion sheet'), of the salient criteria when the task is set but beyond that there is no necessity for them to be tied strictly to particular classes of tasks described in a Work Program or Syllabus.

For purposes of summative assessment, scanning down the columns tells whether a student is generally good at content, organisation, language and presentation, information on which is necessary for mak-ing a criteria-based decision on a Level of Achievement. At any particular time, of course, such a vertical scan will reveal the student's area of greatest weakness, and the area of greatest strength. Note that with this achievement table, both numerical marks and the 'weighting' of each instrument, which is so important in the Radford system, is replaced by an entirely new notion, that of 'weighting' the various criteria according to their importance. It may be, for instance, that the syllabus writers want organisation to count more than presentation. The would have to be stated in the syllabus documents, and reflected in the commentaries attached to the exemplar sets.

A more elaborate teachers' assessment table, in which the Xs are replaced by appropriate Level 3 criteria, could help teachers keep track of the sub-criteria as well. As a planning device, it would signal the fact that opportunities should be created to assess students on most of the Level 3 criteria, enabling the teacher to see at a glance whether data on all the criteria in the syllabus are being gathered through the assessment program. Such a table would provide an effective way of keeping the criteria visible. Scanning across each row, the Level 3 criteria that are salient to each assessment instrument would be immediately apparent. (In English, such a table would replace what is known in the literature as a Table of Specifications or Test Blueprint.) Throwing the focus on the criteria rather than on tasks or instruments could lead to rationalisation in an assessment program, reducing the tendency towards overtesting.

## Conclusion

The comments and suggestions above are intended to help set up a framework for thinking through what criteria-based assessment might look like both conceptually and practically, using English as an example. Although alternative formulations are no doubt possible, the hierarchical structure serves to help systematise criteria (which otherwise have a tendency to float around in list form), and the cross-tabulation of criteria with assessment instruments shows how the criteria could be incorporated into a recording system.

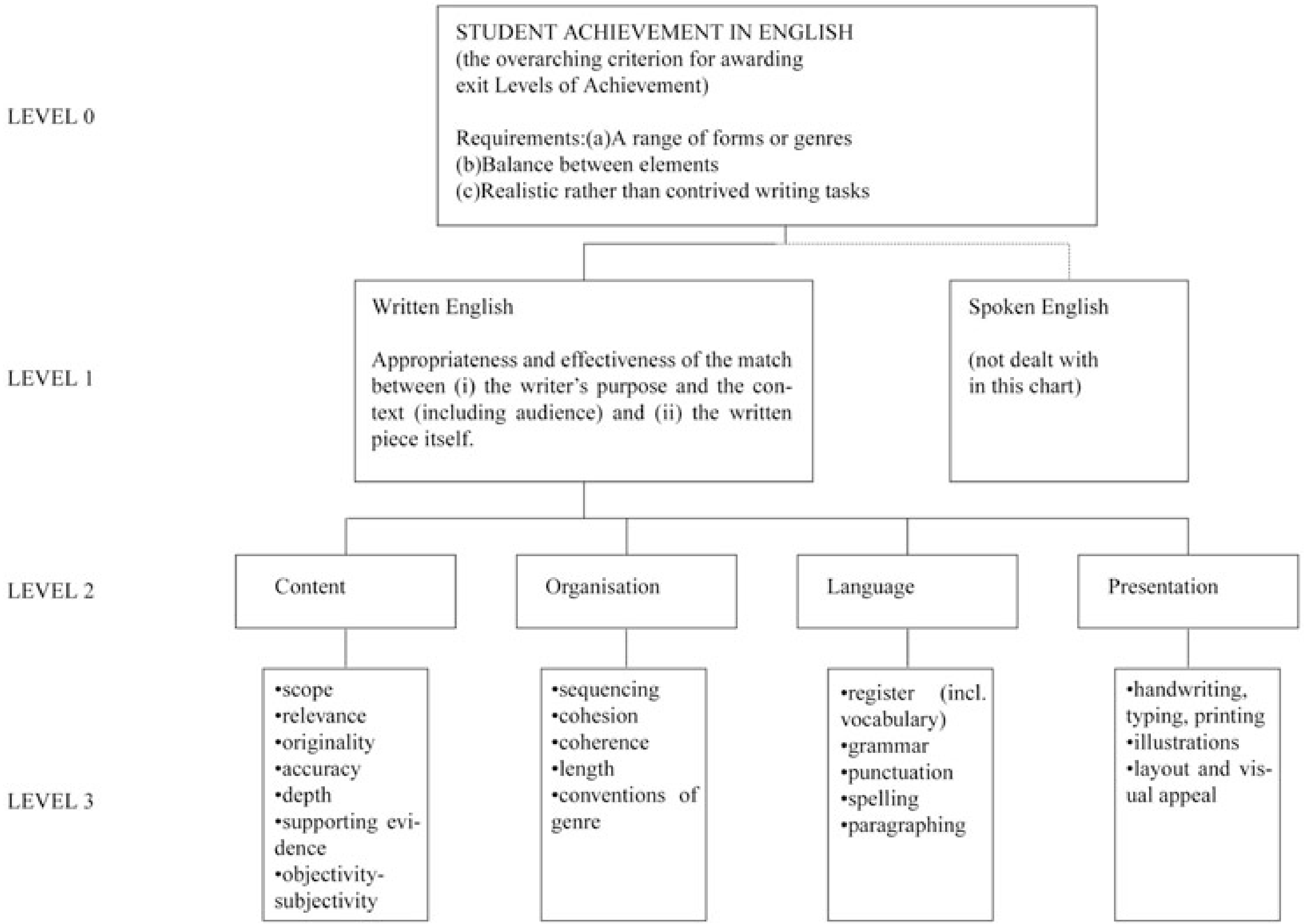



Figure 1: Chart displaying a possible organisation of criteria in English

# Affective Objectives under ROSBA

## Discussion Paper 10

**Abstract**: Affective objectives have to do with interests, attitudes, and values, and constitute an important aspect of education. They have implications for teaching, learning, the curriculum, and the organisation of schooling. Whether they should be assessed, either at all or in specified areas, is an issue worthy of some discussion. In this Paper, it is argued that affective responses should not be incorporated into assessments of achievement.

**Author**: Royce Sadler, Assessment Unit, April 1986

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



The aim of schooling is to help students become more than fact banks and intellectual robots. Education should aim to promote, in Broudy's (1979) delightful words, "clear thinking, enlightened cherishing, and humane judgment". To acknowledge this is to accept the proposition that education is more than training. Exactly wherein the difference lies, and what to do about the difference once it is identified, is a matter that requires some analysis.

Knowledge is, of course, important. But equally important is the student's interaction with the knowledge acquired, particularly how the student approaches, and responds to, the implications of that knowledge. Such interactions draw curriculum workers and teachers into what is called the affective domain, the world of interests, attitudes, and values. It is widely known that achievements in the cognitive domain do not necessarily correlate well with developments in the affective domain, although it is reasonable to expect that, on the whole, students who find a subject interesting are more likely to try hard at it. However, some students who are diligent and intensely interested in a subject may not have what it takes to achieve well in it. Conversely, some students who achieve well adopt a strictly utilitarian approach and are left quite unaffected by all attempts to generate interest in the subject. For other students, the effect may even be negative: all initial interest in a branch of knowledge or the culture may be driven out of them by the way the subject is approached an taught. There is, therefore, no strong causal link between cognitive achievement and affective development, in either direction. That is why the affective domain is worthy of attention in its own right, and why syllabus writers incorporate affective objectives into their documents. But what role do, or can, these objectives serve? Should students be assessed on affective outcomes? The aim in this Discussion Paper is to explore answers to questions like these.

An obvious place to start is with statements of affective objectives in syllabuses. Consider the following collection of typical affective objectives. These have been extracted from a number of current *Senior* syllabuses, and are set out under a simple classification according to type.

**(a) Attitudes towards the subject, or towards schooling.**

- The student develops an interest in and a sensitivity to the possibilities of language (Senior English, 1981).

- The student should experience the artistic reward of participating in solo and group presentation (Senior Theatre, 1981).
- The student will appreciate elegance in mathematical proof, and recognise imagination as a component of mathematical thought (Senior Mathematics, Units 3 and 7, 1983).
- Students will develop a willingness to value the procedures of the historian (Senior Ancient History, Unit 3, 1981).

**(b) Personal qualities.**

- The student should be persistent in efforts in undertaking and completing assigned investigations and work (Senior Agriculture and Animal Production, 1981).
- Students should have the opportunity to value themselves, their personal integrity, and their humanity (Senior Physics, 1984).
- The student should develop a respect for the maintenance and cleanliness of equipment and materials (Senior Graphics, Trial/Pilot, 1984).
- The student should develop a sense of commitment and personal integrity (Senior Speech and Drama, 1981).
- The student should develop a predisposition to proceed with an experiment in a systematic fashion and with purpose (Senior Multi-Strand Science, 1984).

**(c) Relationships with others.**

- The student should have an opportunity to value working co-operatively with others (Senior Multi-Strand Science, 1984).
- The student should demonstrate willingness to cooperate in small groups and class situations especially in the field (Senior Geography, Unit 1, 1981).
- The student assumes increasing responsibility for the collaborative planning of learning (Senior English, 1981).

**(d) Commitment to certain socially accepted values.**

- The student should have an opportunity to develop a personal position with respect to the use of non-renewable resources (Senior Multi-Strand Science, 1984).
- The student should have the opportunity to be objective and honest in representing observations, be ready to admit error and be tolerant of the errors of others, and be receptive of the limitations of man and science (Senior Chemistry, 1983).
- The learner should come to appreciate the achievements of a culture other than his or her own, and see his or her own language and culture in a new perspective (Senior French, 1984).
- The student should have the opportunity to become aware of the risk situations and willingly conform to safety regulations (Senior Biological Science, 1983).
- The student will develop aesthetic attitudes and appreciations, and be flexible and tolerant to all forms of visual art (Senior Art, 1984).

**(e) Responsibilities in adulthood.**

- The student will exhibit a desire to participate in physical activity as a vital component of a healthy life style (Senior Health and Physical Education, Element 1, 1985).
- [The student should] develop a commitment to the application of economic principles in solving personal economic problems and in exercising influence as a citizen and voter (Senior Economics, 1984).

- Students will [show] a willingness to develop desirable attitudes towards their responsibilities as Australian citizens including participation in decision-making processes affecting both domestic and external relations (Senior Modern History, Unit 8, 1984).

**(f) Pseudo-affective objectives.**

Some so-called affective objectives, on examination, are actually cognitive objectives in disguise, in spite of the use of terms such as value, appreciate, and be aware of. For example, one of the affective objectives in the chemistry syllabus is that students should “appreciate the role of nuclear chemistry in the quest for energy”. It is clear that appreciate here has a different meaning from what the term has in relation to a student’s personal response to say, a poem or the environment. Other examples are:

- The student will be aware of the existence of professional secretarial bodies (Senior Secretarial Studies, 1982).
- The student should be able to make rational judgments and assessments of alternative courses of action and of the likely consequences, and to approach issues in economic geography which involve value judgments (Senior Geography, Unit VII, 1981).
- The student will appreciate that Modern Algebra involves the manipulation of symbols without implying any interpretation as numbers (Senior Mathematics, Unit 9, 1983).

Such objectives should be rewarded, and transferred into the cognitive (or in some cases psychomotor) objectives for the subjects.

## Assessment of affective objectives

The general educational literature contains many calls for more attention to be paid to affective outcomes in education. A way to achieve this, say some writers on educational assessment, is to treat the affective domain like the other two domains, and require students to be assessed on affective outcomes. The logic of the argument goes as follows: aspects of schooling that are not assessed tend to be neglected; affective outcomes typically are not assessed; therefore affective outcomes tend to be neglected. The implication in the argument as to causality is intentional. This reasoning thus constitutes a case *for* assessing affective outcomes, even though in its more extreme formulation the primary motivation for assessment lies less in the use to which the resulting information could be put than in the instrumental effect such assessment presumably would have. There are, or course, other quite different methods of encouraging students and teachers to value affective outcomes and to take them seriously.
Just as there is an argument *for* the assessment of affective objectives, so there are at least two arguments *against*. The first has to do with the practical difficulties of measurement, and is the one taken up by the writers of the ROSBA Report. They argued that the attainment of affective objectives cannot be measured directly or objectively.

### Recommendation M12 reads:

> Board assessment policies should relate to those dimensions of achievement which can be validly measured either directly or indirectly, viz. process achievement, content achievement and skills achievement. Reports on student acquisition of the non-directly measurable objectives of a syllabus (e.g. the affective objectives) which cannot be made objectively, should be made as appropriate by the school on the school leaving certificate and/or the school progress report.

The reservations of the ROSBA committee have considerable validity, in that attitudes and values are, for a number of reasons, difficult to measure with accuracy. Difficulties include faking, self deception, a tendency to give responses that are socially desirable, and semantic problems in wording items. However there do exist a number of techniques, some of them simple, some of them ingenious and sophisticated, for the measurement of affective responses. Measurement impediments can, except for such adult-role objectives listed under (e), be substantially overcome with enough effort.

The second common argument is that the promotion of particular attitudes and values is not the proper function of the school, especially when schools are funded wholly or in part from public monies. For a school to take a particular line in the affective domain represents, so the argument goes, at best an invasion of privacy, and at worst indoctrination, social engineering, and conditioning. In analysing this argument, it is absolutely essential to differentiate among several classes of values according to their level of support in the community. It is common in some quarters to identify some aspect of affective objectives that can be criticised legitimately, and then to imply that anything associated with attitudes and values has no place in the school. This guilt-by-association argument is possible simply because the term ‘affective’ is used as an umbrella term.

In fact, some value positions enjoy extremely wide (if not universal) approval in Australian society. Examples include cooperation, diligence, honesty, persistence, logical thinking, objectivity in reporting, rationality, nonviolent solutions to conflict situation, individual and group safety, non-exploitation of the weak, and preservation of society and the environment generally.

For other issues, there is approval from certain sections of society and disapproval from others. Examples include sandmining, Christian morality, capitalism, nuclear disarmament, multi-culturalism, preservation of particular aspects of the environment such as rainforests, reefs, the littoral zone, historic buildings, and endangered species.

Teachers who push a particular stance in the areas suggested by the first class of values are less likely to be accused of indoctrinating their students than are teachers who push the values implicit in the second class, simply because the question of indoctrination rarely arises when acceptance of a particular point of view or value is almost universal in the community. An examination of Senior syllabuses shows that the inspiration for most, if not all, of the affective objectives is drawn essentially from the non-controversial values of Australian society. The objectives listed under the headings (a) to (e) in the first part of this Paper are typical examples.

## Affective objectives and certification

The way is now open for a sharper focus on the matter of assessment in the affective domain. Given that the majority of affective objectives contained in current syllabuses would receive full community support, and that attainments could be measured with the available testing technology, the question remains as to whether schools SHOULD, in principle, be in the business of assessing in the affective domain for purposes of certification. The issue is now tackled from a quite different direction.

The Board, as a public agency, is in the business of certifying *achievement*, (called *competency* in the original ROSBA Report), not value commitments. The term achievement is never construed in speech or in practice in such a way as to include feelings, beliefs, or attitudes. Achievement is associated with things accomplished, a degree of success, or a quality of performance. None of these concepts makes any sense in relation to a student’s disposition or depth of commitment. Showing a willingness to conform to safety regulations in carrying out laboratory work, or being tolerant of the opinions of others, is not what is normally understood by student achievement, although if a teacher’s students become, say, more sensitive to a foreign culture through studying a subject at school, the *teacher* has, in a non-trivial sense, *achieved* something important.

Because of its concern with achievement, the Board is not willing to allow the value commitments of students to be incorporated into Levels of Achievement. If it did not follow this policy, exit Levels of Achievement would be unspecified amalgams of achievements and attitudes. A student who is unable to perform well but is industrious and sincere could, for example, receive the same recognition on a certificate as another student who achieves reasonably well but is lazy and uncooperative. It is of course true that certain dispositions and value commitments (such as interest and diligence in a subject) are ordinarily reflected in the student’s achievement anyway. In an important sense, therefore, certain affective objectives are measured indirectly by measures of achievement, and this is of course right and proper. The fact remains that for all Board and Board-Registered School Subjects, the only criterion for the award of an exit Level of Achievement in a subject is achievement itself.

Of course, the affective aspects of a Board or Board-Registered School Subject may be assessed and reported on the school leaving certificate. In fact, the ROSBA Report recommended that affective developments be so reported, using descriptive statements. The school is not permitted, however, to incorporate such information into an exit Level of Achievement, even on the school certificate. Recommendation M40 from the ROSBA Report says, among other things, that the "exit Levels of Achievement reported on school leaving certificates issued at the end of Years 10 and 12 should indicate the same Level of Achievement as that shown on the Board-issued Junior Certificate or Senior Secondary School Certificate" for all Board and Board-Registered School Subjects studied. By implication, this effectively precludes the situation where, say, a student with high performance in an affective area (which the school may value highly) could receive a higher 'school grade' either than the official Board Level of Achievement, or than that of another student having the same achievements in the cognitive and psychomotor domains.

There is one final matter so far left unresolved. Under ROSBA, Subject Advisory Committees are required to include appropriate affective objectives in syllabuses. This achnowledges the importance of affective objectives in education, and is in no way inconsistent with the Board's primary concern with achievement. But exactly how are affective goals to influence the curriculum? If they are not to be formally assessed for purposes of certification, can they ever amount to more than window dressing or pious hope? The answer is a clear affirmative for two reasons. First, affective objectives are important in curriculum evaluation, because the worth of a curriculum should be judged by the extent to which it acheives its full educative, and not just its cognitive and psychomotor, aspirations. Second, affective objectives are of critical importance in the selection and organisation of learning experiences. They give an orientation to teaching, and in many cases can help teachers in the choice of one set of learning experiences over another, especially when the cognitive outcomes are likely to be identical. To put it another way, other things being equal one set of activities or experiences is educationally more valuable than another if it is more in keeping with, and more likely to lead to, the realisation of the affective objectives. This is probably the most important function of affective objectives in the curriculum. Among other things, such a perspective makes the earlier argument for the assessment of affective responses irrelevant.

## Conclusion

Affective objectives typically enjoy an uncertain status both in syllabuses and in the curriculum generally. This is due partly to a lack of understanding as to what constitutes a genuine excursion into the affective domain, partly to reservations about the extent to which affective objectives should be consciously pursued, and partly to a belief that they are too personal or subjective to do much about. In this Paper, it has been argued that the primary role of affective objectives is in providing guidelines for the development of learning experiences, and that any record of student achievement has to be literally what it says: no more, and no less, than a record of *achievement*.

## Reference

Broudy, H. S. (1979). Tacit knowing as a rationale for liberal education. *Teachers College Record, 80*, 446-462.

# School-based Assessment and School Autonomy

## Discussion Paper 11

**Abstract**: School-based assessment in Queensland means that teachers have responsibility for constructing and administering assessment instruments, and for appraising student work. But because certificates are issued from a central authority, the assessments must be comparable from school to school. In addition to being school-based, the ROSBA system is criteria-based as well. It is argued in this Paper that using uniform criteria and standards across the state allows for variety of approach in assessment and helps to achieve comparability without destroying the autonomy of the school.

**Author**: Royce Sadler, Assessment Unit, April 1986

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



Compared with external examinations, school-based assessment delegates a number of responsibilities to schools with respect to the summative evaluation of students' achievements. The domains and the degrees of autonomy actually accorded to the schools must be resolved under such a system. It is obvious that there cannot be absolute autonomy in any situation where there is a requirement for comparability, such as occurs when certificates are issued by a central agency (which in Queensland is the Board of Secondary School Studies). Schools therefore enjoy what might be called a 'bounded autonomy'. It is necessary to make the boundaries clear if public confidence is to be maintained, and if schools are to operate efficiently (and feel comfortable) in the system. Where should the line be drawn between complete freedom for the school, and complete control by the central authority? In this Discussion Paper, some of these issues are explored. Comparability of exit Levels of Achievement is completely consistent with the philosophy of ROSBA (and with the spirit of Radford before it). To hold that 'school-based' means that schools have sovereign rights to determine exit Levels of Achievement according to criteria and standards which they set themselves is to extrapolate beyond the intentions of both the Radford and ROSBA committees. Neither of their reports implied freedom to that extent, and both outlined a set of measures to ensure that standards would be the same across the state.

At the outset, however, it is as well to point out that substantial freedoms exist whether the certification system employs external examinations, a mix of external examinations and school assessments, or school assessments alone. Under any of these systems, schools normally have certain *curriculum* freedoms, including the freedom to select content specializations (in some subjects), and to design appropriate learning experiences (in all subjects).

With respect to *assessment*, schools typically are free to decide on the

(a) nature and quality of feedback to students,

(b) total time and energy devoted to assessment during the course,

(c) frequency and timing of assessment activities,

(d) form and manner of reporting to parents,

(e) items, tests, and other instruments to be used for purposes other than final certification.

None of these can be 'claimed' as a benefit of school-based assessment as such. The freedoms available under school-based systems lie elsewhere. There are three distinctive features of school-based assessment, which apply whether the basis for reporting student achievement is norm-referenced or criterion-referenced. The first feature radically extends point (e) above to the extent that teachers themselves make the *summative* judgments about the extent of students' knowledge and the quality of their work *for purposes of certification*. To do this, teachers design and administer their own data-collection programs. The second feature is a consequence of the first: schools are free to decide on the balance between formative and summative assessment. The third feature is that school-based assessment relies on a distributed professional responsibility for operating a system of checks and balances in order to achieve comparability. (In Queensland, this involves officers of the Authority and experienced teachers who sit on District and State Review Panels. The administrative structures for this system are normally under the control of the same central authority as issues the certificates.

If, in addition to being school-based, the assessment is to be criteria-based, both the assessment instruments and the teachers' judgments must be related to the criteria and standards. A strong argument can be mounted for these criteria and standards, in the interest of comparability, to be laid down in the syllabuses. Although this latter requirement places some constraints on schools, there remains considerable latitude in the ways assessment programs can be designed and still be in accordance with specified criteria and standards. It should be obvious, however, that if a criteria-based assessment system is to realise its intentions, the principles and practice of assessment using criteria and standards must be thoroughly understood by the users.

In the operation of a criteria-based certification system, a useful analogy can be drawn with the practice of calling tenders for a product or job. The client makes available the specifications for what is required, and suppliers examine how they can fulfil the requirements. The client is typically less interested in exactly how the supplier proposes to meet the minimum requirements than that the requirements be met somehow. This difference between process (means) and product (end) is, as it turns out, also reflected in standard tendering practice. Specification writers call the former 'descriptive specifications' and the latter 'performance specifications'. Where facilities are available for checking the quality of the end product, performance specifications are generally preferred because they give the supplier of goods or services the greatest possible latitude in selecting a method (or devising a new one) to fulfil the requirements.

Tendering practice provides a perspective from which to view ROSBA. The Board lays down certain specifications (criteria and standards) for the various Levels of Achievement in each subject. Each school examines its clientele, its philosophy, and its human and capital resources, and devises a scheme whereby it can meet the Board's specifications with a certain form of 'product', namely, student achievements. As part of the reviewing process, a school tenders its proposal to a Review Panel, which then accepts, suggests modification to, or in rare cases rejects, the proposal as being adequate for satisfying the specifications for purposes of certification. Unlike the commercial analogy, ALL tenders that satisfy the requirements are, of course, accepted. In some subjects, the 'products' offered may look superficially quite different from school to school, because of the nature of the subject. In other subject, the constitutive elements of the subject, a long tradition of certain forms of assessment, or almost universal use of a particular textbook, may result in essentially similar products. Naturally, greater variety can be expected in some subjects than in others. In either case, the processes the school uses in helping its students satisfy the standards are of only incidental interest at the point of certification.

From a school's point of view, meeting the standards at exit includes:

(a) organizing learning experiences in such a way that students are placed in a position where they can achieve to the limit of their potential;

(b setting interim or progressive targets along the way, semester by semester, so that students have realistic short-term goals that are achievable with the resources available; and

(c) tendering, and making the case as to how the end-of-course achievements of students at the school match with the standards laid down in the syllabus, that is, making the case for ‘equivalent to (in quality) but different from (in detail)’ what is required for the award of the various Levels of Achievement.

The logistics of getting such a system going in the first place are, as it turns out, quite complicated. The criteria and standards contained in the syllabuses apply to the work of students at the end of the two-year courses, but the schools have to receive approval (or accreditation) on the basis of a proposal for assessment submitted at the beginning of the course. At that point, schools have no concrete examples of student work to display, only prospective achievements. Once the initial hurdles are overcome, however, both schools and the Board can refine the processes and develop confidence in using criteria-based assessment.

## Stimulating variety

A central authority has great potential for facilitating change, especially when it not only advocates that distinctiveness among schools is educationally desirable, but also shows how it can be accomplished. Ordinarily it would be insufficient merely to provide the conditions in which differentiation would be possible if at the same time schools were left feeling somewhat insecure about what would be acceptable. Ironic though it may appear at first, a central agency by adopting proper strategies can actively promote change and variety, and still fulfil its obligations to society for comparability. The essence of the ROSBA system is that the responsibility for the specifications as to quality of the end product remains with the Board, but that schools formulate their own methods of meeting the specifications.

A difficulty at present is that many teachers lack the confidence or the skill to prepare ‘equivalent to but different from’ tenders. One way to attack this problem would be for the Subject Advisory Committees to formulate clear criteria and standards, and then to compile (say four or five) different sets of exemplars which are accepted as legitimate ‘realizations’ of the standards on the criteria. In other words, the Subject Advisory Committees could take on board the task of assembling and disseminating multiple exemplar sets that indicate not only the scope of variation possible, but also make it easier for the teaching profession and the public to appreciate the validity of different approaches. An important practical spin-off could well be that the criteria and standards come alive in the presence of difference exemplars. Otherwise there could be a tendency for them to remain as abstractions, on which teachers could all say they agree, while understanding them differently!

It is argued in the next section that exemplar sets would be meaningless without details of the associated criteria and standards to hold them all together, and that the use of statewide criteria does not have to result in tight prescription and bland uniformity. The combination of clear criteria and standards, accompanied by examples of diversity, could do more to encourage heterogeneity in syllabus implementation, local (school-level) specialization in approach, and experimentation with novel forms of assessment than exhortations for schools to branch out. At the same time, it could help fulfil the conditions for equivalent standards across the state.

## The mediating role of criteria

The ROSBA system places great value on both variety and comparability, factors which may on the surface appear to pull in opposite directions. In this section, it is shown how the appropriate resolution lies not in some sort of pragmatic compromise, but in a proper understanding of the crucial role that can be played by carefully formulated criteria. The argument itself is an indirect one, and begins with the concept of comparability. What does it mean to say that things are *comparable*? Actually, there are two easily distinguished meanings recognised by compilers of dictionaries.

The first meaning, which is “able to be compared” with the emphasis on the *able to*, follows directly from the etymology. It is used when there is some question as to whether it is possible to compare two things that are not identical, and is written here as compare-able. The second dictionary meaning is “more or less equal”, and when used this way, the word is written here as com-parable. The proposition about to be argued is that compare-ability of two potentially equivalent things is logically prior to a judgment about the equivalence itself. (In the present context, the “things” are student achievements from different schools.) Compare-ability is essentially *prospective*, and com-parability *descriptive*, of the drawing of a conclusion as to similarity or equivalence. Criteria, it is argued, play a decisive role in making such judgments.

First consider compare-ability. Some things so obviously belong to the same genre, and are so obviously compare-able, that the matter is not given a second thought. For example, performances of a set piece of music by different students are clearly compare-able. So are the acrylic bowls made by students in the manual arts workshop. The relative merits of things from within a particular genre can often be decided without resort to explicit criteria. In other words, the criteria are implied by the context, although they have to be brought to the surface and made explicit if teachers’ qualitative judgments are to be substantiated to students, other teachers, and parents.

The question of compare-ability for things from different genres (such as a watercolour and an oil painting in art, or a cartoon and a poster in English) is more complicated because the things may be compare-able with respect to some features or aspects and not with respect to others. For example, an egg and a glass of milk are not compare-able with respect to shape (compare-ability in this case is a nonsense idea), but they *are* compare-able with respect to nutritional value of cholesterol content. Here, a criterion by which to assess compare-ability is not immediately suggested by the context and has to be stated explicitly. Once it is stated, the compare-ability issue is settled and it becomes sensible to ask whether the things might also be com-parable.

The need for criteria which cut across several (but not necessarily all) genres is directly dependent on the degree to which surface dissimilarities may obscure the deeper commonalities. But it is precisely these commonalities (which are signalled by the criteria and standards in a subject) that help to characterise art as art, and English as English. Under school-based assessment, schools make choices with respect to such aspects as content and assessment instrumentation, within the general framework of a syllabus. As a consequence, subjects are ‘expressed’ differently in different schools. Uniform criteria (and standards), by providing common links between these different expressions, make compare-ability possible.

There are, of course, things that belong to genres that are apparently so distant from one another that no common aspect springs to mind at all. For example, most people would say that a banana and a truck are simply not compare-able (although even here size is a possible criterion). The ROSBA system can be thought of as pitched at an intermediate level: neither complete uniformity, nor bananas-and-trucks, but different expressions of the same ‘essence’ of a subject, in which a common set of criteria and standards can be instrumental in moving towards first compare-ability, and then com-parability. That is why it is so important under ROSBA not only to have criteria, but for all school to work to a *common* set of criteria which, therefore, must be stated in syllabuses.

## Conclusion

Under ROSBA, schools enjoy certain freedoms as to how they assess their students, but they must conform to the requirement that the tasks set for students produce outputs that are compare-able with the outputs from other schools. This means that a school cannot be free to decide its own criteria, but must ensure that the criteria contained in a syllabus can be applied to the achievements of its students in that subject. Furthermore, in the interests of accountability and public confidence, work associated with each of the exit Levels of Achievement must be com-parable in the descriptive sense. That is, the standard of work for a particular Level of Achievement in a subject has to be, within the limits of human judgment, the same from school to school.

# Defining and Achieving Comparability of Assessments

## Discussion Paper 12

**Abstract**: Four interpretations of comparability are identified and discussed in this Paper: comparability among subjects, students, classes, and schools. The last of these, comparability among schools in each subject, is a crucial concern in a school-based assessment system in which certificates are issued by a central authority. Only when the Levels of Achievement have consistent meaning across the state can public confidence in the certificate be maintained. It is argued here that achievement of comparability is fully compatible with the concept of teachers as professionals, and with accountability of the profession to the public at large.

**Author**: Royce Sadler, Assessment Unit, January 1987

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



It is convenient for the purposes of this Paper to distinguish four different faces of comparability, which are discussed in the following order: comparability among subjects, comparability among students within a class, comparability among classes (with different teachers) within a school, and comparability within a subject among the schools which make up an education system. Although the underlying interest is in the particular form of school-based assessment which uses criteria and standards for grading student achievement, comments about other systems and grading principles are introduced at various points in order to clarify the issues. The topic of comparability is treated in general terms first. Towards the end of the Paper, a number of implications for assessment under the ROSBA system in Queensland are spelled out.

## Comparability Among Students

It might be asked whether a grade of “A” (or 7, or Very High Achievement) in one subject, say music, represents an accomplishment that is equivalent in some sense to what is required for the award of the same grade in another subject (say, English) at a particular level of schooling. Comparability of this type is important when the grades awarded in different subjects are treated as though they were interchangeable, with one subject being regarded as equal to any other for a particular purpose. If, for example, school leavers require a certain number of As, Bs, or Cs for entry into employment or further education, it would not do for As to be easier to achieve in some subjects than in others.

One technique often used for investigating among-subjects comparability is fairly straightforward: for all students in a given population who study, say, both music and English, the proportions receiving each of the different grade levels are compared. If the distributions of grades are similar, the subjects are deemed to be comparable. In national surveys such as those carried out in Britian, the assumption is made that ove a whole school system, local variations balance out and the similarity of grade distributions provides a valid measure of the comparability of different subjects.

It is possible, of course, to *ensure* (rather than just check after the event) similarity of grade distributions across subjects, in one subject across successive years, or across both subjects and years, by adopting a policy of grading on the curve, as under the Radford scheme. The exact shape of the curve is of no consequence; it could be Gaussian (normal), rectangular, or otherwise. Predetermined proportions of students are awarded the various grade levels. If one is willing to assume (a) that the (overlapping) populations of students are equally able academically, or (b) that the average performance of students in a subject remains constant from year to year, it can be argued that grading on the curse achieves comparability respectively (a) among subjects, or (b) over years. In practice, these conditions are fairly stringent and are almost never met in systems where students self-select into subject and courses, and where the proportion of the age group staying on at school past the official school-leaving age (to end of secondary education) steadily increases (as at present) or decreases.

A second form of among-subjects comparability arises when assessment in different subjects are combined to form aggregates, which are then used to compile on order-of-merit list for purposes of selecting students for higher education or awarding scholarships. Whereas the grading scale usually has between four and nine levels, achievements are typically reported on a finer scale for the order-or-merit list. Subject scores are then adjusted using either an external moderating test (which every student takes), or an iterative (internal) scaling procedure. In Queensland, statistical adjustment is carried out using the results from the Australian Scholastic Aptitude Test to align each subject against the others (prior to aggregation and a second stage of scaling) for the compilation of the Tertiary Entrance Score. That such scaling is possible does not necessarily imply, of course, that something similar ought to be attempted for purposes of certification.

Both of the approaches outlined above (equalizing grade distributions, and statistical scaling) closely identify comparability with the level of difficulty of the subject as experienced by students. As a consequence, the comparability that is achieved is on one dimension only, because difficulty level is only one of many possible properties of assessment in a subject. The rationale, of course, is that students should be neither advantaged nor penalized for choosing particular subjects or combinations of subjects.

Beyond the issue of the relative difficulty of subjects lies the intrinsic conceptual complexity of the subject matter, and of the characteristics of the testing program in relation to how the material is presented and learned. It is doubtful, however, whether attempts to untangle the philosophical and practical problems of judging the comparability of what are in reality incommensurables are likely to achieve enough to be worth the effort. Clearly different criteria for judging performance are appropriate for different subjects. In large measure that is why each subject is able to make its own unique contribution to personal development, and why students opt for different subjects according to their aptitudes and interests. Although there is a general expectation that subjects should be pitched at a level appropriate to a particular stage of schooling, it is a moot point whether it is possible to ensure that subjects are “equally rigorous academically”. In some situations such a requirement might result in no awards at all at the highest grade level for subjects which attract less able students. Almost certainly, this would be negatively motivating for those students, in that hardly any would taste of success, for quite arbitrary reasons. Apart from the obviously ulilitarian goals mentioned above, it should be sufficient to allow each subject to go its own way, and leave the interpretation of different achievement grades in different subjects to those who make use of them. Only naive employers would look at Levels of Achievement without regard to the subjects studied.

## Comparability Among Students Within a Class

The second face of comparability is related to the fairness with which a teacher assesses student work. If grades are awarded fairly, the achievements of students who are given the same grade are, within accepted tolerances, comparable. That is, fairness implies that all students are awarded grades that reflect accurately the quality of their work, and that the assessments are not affected by extraneous factors. Potential threats to fairness may be grouped under two headings: (a) a teacher’s personal grading practices, and (b) the design of the assessment program, particularly in the matter of choice in assessment tasks.

There is extensive evidence in the literature of how teachers' assessments can be influenced by such factors as the congruence between teacher's personal values and those of the students, by teachers' knowledge of personality and other characteristics of students (including their diligence, classroom behaviour, and agreeableness), by the order in which teachers assess or grade student work, and by handwriting. Some of these can be minimized by suitable techniques. For example, the seriousness of order effects can be reduced by randomizing before each grading session. Ideological bents and halo effects are more elusive, but their seriousness can be reduced by arranging for cross-marking with other teachers who do not know the students.

Comparability is also threatened when students are given a choice of tasks. Suppose, for example, that students in a particular subject are given a choice of four out of seven items on an examination paper. (Actually it makes no difference to the thrust of this argument whether the examinations are external or school-based, or whether the options are for examination questions, assignment topics, or take-home projects. The issue is the matter of choice among alternatives.) Suppose that the items on a hypothetical examination paper, stripped of their specifics, are a follows:

1. Describe the main characters of A.
2. Outline briefly the principal contribution of B to C.
3. Illustrate the influence which D had on E by giving three different examples.
4. Explain the significance of F for the development of G.
5. Compare the views of H and J with respect to the success of K.
6. Evaluate the proposition that L was the principal or only cause of M.
7. Given that conditions N were in force, estimate the extent of P's contribution to the emergence of Q.

In this reduced form, of course, the examination paper is somewhat cryptic. Good test tasks typically have more detailed specifications. But the example is sufficient to demonstrate the point that a student who elects to respond to items 1, 2, 3 and 5 submits work that is primarily descriptive. A choice of items 3, 4, 6 and 7, however, obviously is more demanding in that it requires higher-level critical skills. If an unrestricted choice is allowed, it becomes difficult to judge the equivalence of descriptive and critical responses from students, and hence to achieve comparability. The situation is actually more complex than this because more than two categories can be identified, but the problem remains so long as students are able to choose among options that do not make cognitively similar demands. One practical solution is to insist that students choose similar proportions of items from the several categories.

## Comparability Among Classes Within a School

Whenever classes in a subject are taught by different teachers, comparability of assessments among classes is an important consideration, even when the same tasks or instruments are used. One teacher may teach to the test, deliberately or inadvertently. Another may prime students to give superb but atypical or artificial performances. Yet another may be overzealous in the opposite direction, even to the point of not preparing students adequately for the test. In addition, the wording in some tests may be more familiar to students in one class than in another, not by design, but because teachers who are thoroughly conversant with a body of subject matter are sometimes unaware that they teach using a particular, and to some extent idiosyncratic, vocabulary. The teacher setting the test may have pet phrases or forms of expression which are semantically equivalent to the forms of expression used by teachers of other classes, but are nevertheless unfamiliar to those classes. Or the teacher setting the test may run so close to production deadlines that other teachers have too little time to review the test thoroughly. It is obvious that these matters can be dealt with only within the context of the school, and that schools should develop policies that encourage internal moderation, openness, and consensus with respect to assessment so that the probability of fair assessments for all classes is high.

## Comparability Among Schools Within an Education System

This is the crucial issue for a school-based assessment system where certificates are issued by a central authority. It is not an issue when there is a program of full institutional or course accreditation, or (*validation* as it is called in the United Kingdom), that is, when an education authority accredits schools or colleges as being competent to design and teach a curriculum, assess student achievement, and issue certificates in their own right. School-based assessment, by definition, stands for non-uniform assessment programs and instruments from school to school. It is therefore necessary for this aspect of comparability to be opened up and discussed. In particular, what comparability means and implies for practice has to be explored, so that appropriate steps can be taken to try to achieve it.

Within a context of statewide certificates and competitive entry to tertiary education through a central admissions bureau, any school-based assessment system will stand or fall according to whether there appears to be comparability among schools in a particular subject. The importance of comparability as an issue is directly proportional to the consequences of a wrong or unjust decision. In the present social climate of reduced employment opportunities for school leavers, and with the demand for places in tertiary institutions running at several times the supply, competition is intense. Parent naturally feel extremely sensitive to the comparability issue, and want their students to compete on an equal footing with other students from the same school, and with students from other schools. Given these considerations, it is clear that care must be taken to set up and maintain procedures that are soundly based and that foster public confidence.

A certificate issued by a central education authority has a significant social value in that it acts as a leveller, recognizing the achievements of students no matter whether they study at a large metropolitan college with high visibility, or at the only secondary school in a country town. The imprimatur of the issuing authority give the certificate statewide currency, and protects the interests of students who attend small, isolated, or little-known schools. It is therefore an important factor in achieving equity in education. Wherever a centrally authorized certificate is issued, no credible case can be made for a variation in standards from school to school or from district to district to take into account what might be called “local conditions”.

Comparability of achievements among schools does not attract much public attention under external examination systems, of which there has been a long tradition both in Australia and overseas. Presumably, the assumption is that when the contents of the examination paper are kept secure until the scheduled date, and when all students sit for the same examination paper under rigidly controlled conditions, all students have equal opportunity to demonstrate their achievements. But although the public may seem generally satisfied that all students face the same challenge, comparability is always a latent issue. One of the reservations about the comparability of external examination results arises from the recognition that not all students are equally suited to the formality associated with external examinations, and as a result do not perform as well as they should. In addition, anomalies sometimes occur through the ways teachers prepare students for the examinations. For example, chief examiners have been known to comment about the surprising uniformity of scripts from particular schools in certain years. Sometimes the performances are uniformly on target, sometimes uniformly wide of the mark. Such a phenomenon occurs when teachers try to tip examination questions, and coach or otherwise channel their students’ efforts into producing model answers to the question expects. Other reservations have to do with choice of questions (a matter touched upon above), the consistency of markers (both over time, and from one marker to another), and the comparability of grades in a subject from one year to another. But by and large, external examinations are often able to provide a degree of comparability which teachers and the public find satisfactory.

In some states and countries, hybrid systems have been developed. Information derived from two sources, one external to the school and the other internal, is used to create a composite measure of achievement for reporting on the official certificate. The external source typically consists of either public examinations, or reference tests based on core material in the curriculum. Either the results from these tests are used directly as a formal component of the composite grade for each student, or group statistics are used to scale school assessments. In the latter case, an individual student's performance on the external test does not have a direct influence on that student's final grade in the subject.

To reiterate, comparability among schools is a live and important issue whenever (a) the assessment is school-based, and (b) the achievement certificate is issued centrally. When students are to be assessed against specified criteria and standards (rather than against one another), such a system raises conceptual and administrative problems of a unique kind, and requires considerable ingenuity in the development of acceptable and effective procedures. Furthermore, in discussing the question of comparability, it is important not to get confused between the ultimate intention, and the mechanisms used to try to achieve comparability. One mechanism, which is here called Model A, is to let schools develop their own criteria and standards, and submit them to the certificate-issuing authority for review, for adjustment if necessary, and ultimately for ratification. Assuming that it is just as serious an error for a school to set its standards too high as it is to set them too low, the business of deciding whether standards developed within the school are acceptable for purposes of certification is necessarily labour-intensive. Furthermore, if the authority's expectations are not made explicit (and therefore accessible to the schools), the schools may be set off on a course of stressful trial-and-error learning that may sap the morale of the teaching staff.

An alternative (and more efficient) mechanism, here called Model B, is to develop statewide criteria and standards and incorporate them into syllabus documentation. Student work can then be appraised at the point of exit in relation to clearly specified standards which are binding across the state as a whole. Because all teachers use the same standards as benchmarks, comparability among schools is likely to be achieved. Moreover, the process of achieving a working consensus on standards is simplified. In any case, the principle of having standards that are explicit and accessible applies as much to the authority (for the benefit of schools) as it does to schools (for the benefit of teachers and students).

The concept of "equivalent to but different from" is one that needs further teasing out. Comparability among schools means that the performances of all students who are awarded a particular grade in a subject are of equivalent quality, regardless of the school attended. This does NOT imply that students should tackle identical tasks in all schools, nor that test instruments and assessment programs should be similar in format, or standardized in some other way. It does involve comparing things that are commensurable, for the reason that the "things" are achievements within single subject. It is natural to ask how student achievements that are measured by means of different assessment programs, and sometimes on different content, can be considered equivalent. Perhaps an analogy might help to make it clear that this can, at least in principle, be carried out. Real estate valuers are adept at arriving at a figure for the "value" of, say, a residential property. Houses can differ in every conceivable way, and land parcels differ in area, aspect, location, and amenity. Yet a valuer is trained to assign a dollar value, say $90,000, to a particular piece of property. Other properties worth $90,000 may be quite different in composition, but a good valuer is able, within a fairly small margin or error, to estimate market price accurately.

What is asked of teachers under school-based assessment is that their professional judgments be refined to the stage where the teachers can appraise "equivalent to but different from" performances in much the same way a real estate valuer does. And just as the valuer's figure can be authenticated if necessary by placing a property on the market, so a school's grades can be authenticated by reference to the authority's criteria and standards.

## Comparability and Accountability

How the public perceives the final result is important for any system of school-based assessment. It is not sufficient to argue that all teachers, by virtue of their training, are professional persons and that therefore their judgments cannot be called into question. Procedures must be set up so that justice is done, and is seen to be done, for all students. Making assessments which, although school-based, are comparable with those made in other schools is an important part of the professional activity of teachers. Similarly, accepting responsibility for the comparability of those assessments is part of a teachers' professional accountability.

Accountability is, of course, an accepted element in most service-oriented occupations. In some professions, *independent* checks on professional judgements exist. For example, if a solicitors's legal procedures do not result in properly prepared documents for a conveyancing transaction, a change of ownership cannot be registered at the Titles Office. A medical practitioner whose diagnoses are frequently in error may find continuing in practice difficult. An engineer or architect who designs a building or bridge-which subsequently collapses is likely to have a short career. Many pathology laboratories carry out, voluntarily and at their own expense, checks of their own procedures. One way in which they do this is to have sent to them, from a central laboratory, biological samples whose characteristics are to be determined. (The Quality Assurance Program operated by the Royal College of Pathologists of Australasia, for example, exists solely for the purpose of providing self-monitoring within the profession.) Biological laboratories, by keeping a constant check on their accuracy, are thus able to keep themselves "calibrated".

In order for a certifying agency to be able to organise and give public assurances about comparability, it must have access to assessments from different schools, subjects, and students. Some object to this principle on the grounds that it is intrusive, and denies the professionalism of teachers. Most people, however, would acknowledge that not all teachers want to play the assessment game completely fairly. It is common knowledge that in any school system, a small proportion of teachers seek to advantage their own students (unfairly) if at all possible. Teachers and schools who do play straight have little to fear from independent checks. Most teachers welcome it in the interests of public accountability provided that the checks are appropriate, respect the professionalism of teachers, and make use of the special knowledge teachers have of their students.

## Comparability and ROSBA in Queensland

From the first introduction of school-based assessment, comparability has been an important goal. The Radford report refers to comparability of assessment (section 6.14), and the ROSBA report has a section on the maintenance of statewide standards in the interests of public accountability (section 7.01). It is also featured in some of the leaflets and bulletins produced by the Board of Secondary School Studies for the information of schools, parents, employers, and Review Panels. For example, the leaflet entitled "The Senior Certificate (New Format)" contains the following:

## What about comparability?

To assist teachers establish standards across the state, panels of teachers meet to examine school assessment instruments and students' work. Teachers on the panel offer advice to schools concerning standards of assessment. In this way, acceptable standards of assessment are achieved throughout Queensland.

Comparability is implicit in all of the procedures set up for accreditation, monitoring, and reviewing under ROSBA. The intention is clearly that parents and employers should be able to have confidence in any certificate issued under the authority of the Board. As is well known, ROSBA is being implemented over a period of time. Although all schools are technically on stream, at the time of writing it cannot be said that ROSBA is fully developed. Future experience with school-based assessment will reveal further need for refinement and adjustments, and until these are carried out, it would be premature to judge the success of criteria-based assessment. To date, the Board has no concrete research evidence whether or not comparability is being achieved, but in, say, five years' time it would be appropriate to mount a research study to investigate this aspect. In the interim, it makes sense to continue to identify directions and develop processes which, in theory at least, hold good prospects for ensuring that comparability will be achieved. Substantial progress has already been made in encouraging schools to provide a variety of curriculum experiences, including experimental and non-standard ones, by which students may cover what is in the syllabus. At the same time, all Work Programs have had to meet the general requirements set out in the R2 Form before being accredited. Among other things, schools have been developing their own criteria and standards and submitting them to Review Panels in the manner associated with Model A described earlier in this Paper. It remains to complement this work with further developments in the comparability of exit Levels of Achievement.

In practice it would be conceptually and administratively fairly simple to do this by shifting progressively toward Model B. This would make the comparisons of Work Programs with syllabuses, and the comparisons of student work with Work Programs, of lesser importance in the achievement of comparability. Such comparisons would still be of considerable significance for curriculum development in schools, but they would be less crucial in trying to attain statewide comparability in exit levels for purposes of certification. The Boards responsibility would be primarily to ensure that the external evdence of the quality of a student's performance corresponds with the criteria and standards specified for the award of the various exit Levels of Achievement. That is, there would be a stronger focus on comparability of assessments as an end result, than on the *processes* introduced to facilitate it. To implement Model B, the Board would work steadily towards the production of criteria and standards in such a way that they

(a) Are explicit and accessible to schools;

(b) represent, so far as is possible, consensus across the state;

(c) are developed cooperatively, using the resources and experience of Syllabus Advisory Committees, Review Panels, and classroom teachers; and

(d) steer a middle course between on the one hand tight specificity (with its attendant narrow conformity), and on the other hand loose ambiguity (which would threaten the achievement of comparability).

If Model B were to prove fully effective, it would streamline the monitoring and reviewing procedures, making them less labour-intensive for teachers and Review Panels. Ambiguities as to what the five exit Levels of Achievement should mean in a particular subject would be significantly reduced, making it easier for teachers to classify student performances, and decreasing the probability of errors of judgment. Model B requires convergence on criteria and standards, but this is compatible with divergence on ways to achieve and measure them. (See Discussion Paper 11, *School-based Assessment and School Autonomy*, for a fuller discussion of this aspect.) Indeed, teachers would have a high degree of flexibility in designing assessment programs, especially with respect to (a) the achievement of a balance between formative and summative assessment, (b) the format and style of items and assessment instruments, (c) the nature and quality of feedback, (d) the total amount of time and energy devoted to assessment as such, and (e) novel or experimental ways of assessing students.

There are, however, several developments that need to take place before Model B can become operational. First, the standards need to be formulated, and in such a way as to allow schools to recognise performances of equivalent quality even when those performances are in a form different from those nominated as typical in the standards specifications. Second, there is a continuing need for in-service seminars or other induction programs so that teachers can gain experience in assessing student work against established criteria and standards. Meanwhile, until a fully standards-based system is in place, it makes sense not to discourage the direct comparison of student work from different schools.

# Towards a Working Model for Criteria and Standards under ROSBA

## Discussion Paper 13

**Abstract**: This paper is concerned with clarifying the use under ROSBA of the terms 'criterion' and 'criteria' and with arguing the case for specifying standards within this nomenclature. A model is then presented of how criteria and standards might ideally operate under ROSBA.

**Author**: Marilyn McMeniman, Assessment Unit, January 1987

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Board of Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. It should not b taken as representing Board policy, and is being circulated within the teaching profession to obtain reactions to its contents.



## Introduction

This paper sets out to do three things:

i. to clarify the use of the terms 'criteria' and 'criterion' in ROSBA syllabuses, Work Programs and literature generally;

ii. to justify the need for specifying standards within this nomenclature and structure; and

iii to demonstrate how to specify such standards.

End-of-course criteria and standards are addressed, but the paper stands back from proposing the sub-structure necessary for awarding the five end-of-course categories of achievement to students who exit after 1 or 2 semesters.

In the process, the paper attempts to bring ROSBA terminology into line with what is accepted internationally in educational literature. It makes respectable the term 'criteria-based assessment' inasmuch as it allows the word 'based' to explain the special Queensland departure from criterion-referencing but justifies the entire term by preserving the meaning of 'criteria'.

## Defining 'criteria'

An examination of documents that have been produced under ROSBA shows there is, as yet, no one common use of the term 'criteria'. Thought one might expect that 'criteria' would be a key concept in a system that is purportedly 'criteria-based', more importance appears to be attached to the adjectives that immediately precede the term 'criteria'. In the original ROSBA Report, the terms 'performance criteria', 'achievement criteria', and 'assessment criteria' are all used within the contexts of P11 and P17 (two of the 36 Major Policy recommendations), and all appear to refer to appraisals of student achievements. In Board syllabuses however, 'assessment criteria' and 'summative assessment criteria' appear to refer to those combinations of student performances which qualify students for the five exit Levels of Achievement. In still other cases, schools have used the term 'detailed criteria' to distinguish their own assessment criteria from the 'syllabus criteria' as laid down in the syllabus. The modern language syllabuses include the term ‘macroskill criteria’ to refer to criteria that relate to the four skills of listening, speaking, reading and writing.

Further, there appears to be some confusion in the interpretation of the term ‘assessment criteria’. In some cases, this is used as a omnibus term to encompass a whole host of aspects of assessment including length of the test paper, time limits for particular questions, scoring keys, types of instruments to be used etc. In such cases, the so-called ‘assessment criteria’ are not really criteria at all, but refer to the test instruments and *conditions* under which student achievements are to be assessed, and not to the criteria by which appraisals of student performances are to be made.

It is hardly surprising, then, that one becomes slightly lost when trying to tease out the precise use of the term ‘criteria’. For the purposes of this paper and in keeping with the *philosophy* of ROSBA, the criteria by which student performances are assessed are those that are implied by the objectives of the course which, in turn, are based on the characteristics of the subject. Although this derivation appears logical enough, these relationships are sometimes ignored in practice. That is, the subject-related criteria, the criteria which are either implicit or explicit in General Objectives, and the criteria used to assess student performance are regarded as three separate sets which are not necessarily directly related to one another. Logically, these criteria should all be more or less congruent, even though they may be expressed in different terms.

For example, the criteria by which appraisals of student performances in Year 12 Music are to be made (cf. Appendix I) are those that feature in the objectives for Music. In addition, the four criteria viz. creative writing skills, practical music-making, aural perception, and musical knowledge represent what the subject is in Queensland High Schools. The exit standards which represent the range of student performances on the criteria embody ‘sub-criteria’ such as (for practical music-making) performance, instrumental or vocal interpretation, sight-reading, improvisation, and conducting. Typically, the assessment instrumentation will need to reflect the different sub-criteria in order that judgements of the position of students along the criterion can be made.

In some subjects, there may be ‘enabling’ criteria that are relevant throughout the formative stages of a developmental course of study. These will be *subsumed* as the students near completion of the course. It may be that these enabling criteria are quite numerous. Many teachers in commenting on an essay, for example, may draw the student’s attention to a number of these smaller-scale, enabling criteria (which may be specific to a task) on which the quality of the piece has been judged. While these may not appear on the ‘exit’ list, they are vital to the return of information on performance to the student at certain points throughout the course of study. Where enabling criteria are used, teachers should take care to point out how these mesh with the criteria used for exit-level assessment.

To summarise, then, the use of the term ‘criteria’ is restricted to those properties, dimensions or characteristics by which student performances are appraised. These characteristics are embodied in the objectives of a course (that is, the intended outcomes of a specific secondary program). The objectives, in their turn, are based on the salient dimensions of the subject. Standards refer to fixed points (or benchmarks) along the criteria, representing qualitative differences in performance.

## Specifying standards

There are at lease two common ways of conceptualising and representing criteria and standards under ROSBA. Both have precedents in the wider literature, and both methods appear to have adherents within the system in Queensland. Method I separates criteria and standards as in the schema in Appendix I, and Method II fuses them together in verbal statements.

### Method I

Essentially, this approach necessitates the nomination of a set of standards on those criteria by which student performances in a particular subject are to be judged on exit from the course. The first task is to identify the criteria. As the classification to Levels of Achievement are made in the great majority of cases at the end of the course, the criteria selected should be those that relate to the end of either a whole course or non-sequential units.

Having identified the criteria, a number of standards on each criterion are nominated to enable differentiation of student performances. For example, using the criterion of problem-solving, student performances could be differentiated at points along a continuum, ranging from those that indicate only very little ability to solve familiar problems, to those that show an ability to solve *novel* problems successfully. In addition to establishing standards which relate to the identified criteria, judgements need to be made on whether performance at a given level (say, a Sound Achievement under ROSBA) constitutes achievement of the ‘expected’ standard on every criterion, or whether, for example, a performance that does not meet expectations on one criterion is acceptable provided it is off-set by performances that surpass expectations on all or some of the others. If this compensatory or trade-off notion is used, a statement of all permissible trade-offs relevant to each Level of Achievement should be part of the overall method or schema.

Apart from the Assessment Unit of the Board of Secondary School Studies (1985), several other bodies concerned with assessment have found it useful to conceptualise and operationalise criteria and standards in this way. Examples to date include:

1. Assessment and Certification of Literacy of Secondary Students, Education Department of Western Australia, 1985.
2. Royal Society of Arts Examination Board, (French, German, ESL), 1985.
3. Royal Australian Air Force. (personnel)
4. Department of Defence (Army). (personnel)
5. Spearritt and Colman (Speaking) ESL Proficiency Chart, 1983.
6. City and Guilds of London Institute, 1981.

All six assessment bodes have developed schemas similar in format to that presented in Appendix I.

## Method II

The second method of representing criteria and standards is somewhat different. Instead of *separating out* the standards and criteria as the schemas above prefer to do, this method focuses directly on the Level of Achievement statements. The usual presentation is that, opposite each of the five ROSBA levels, the criteria and standards are fused together and imbedded in verbal descriptions, and no attempt is made to distinguish between them. An example of this is presented below:

| |
|---|
| **Very High Achievement**<br>The student should demonstrate a detailed knowledge of the content, a very high degree of accuracy in the application of definitions, formulae, and learned procedures, and a high degree of initiative and success in applying knowledge and techniques to unfamiliar situations. |
| **High Achievement**<br>The student should demonstrate a detailed knowledge of the content, a high degree of accuracy in the application of definitions, formulae, and learned procedures, and some initiative and success in applying knowledge and techniques to unfamiliar situations. |
| **Sound Achievement**<br>The student should demonstrate a general knowledge of the content, accuracy in the application of definitions, formulae, and learned procedures, and some initiative in applying knowledge and techniques to unfamiliar situations. |
| **Limited Achievement**<br>The student should demonstrate some ability to recall formulae and definitions, to apply formulae, and should recall and apply learned procedures in routine situations. |
| **Very Limited Achievement**<br>The student has gained some experiences in the course of study. |

Thus, for this subject, the key criteria seem to be: (i) knowledge of content; (ii) application of definitions, formulae and leaned procedures; and (iii) application on knowledge and techniques to unfamiliar situations. The standards are implied in descriptors such as ‘detailed’, ‘general’, ‘high degree of accuracy’, ‘very high degree of accuracy’, ‘recall’, etc. However, because the *standards* are not defined explicitly, it is somewhat difficult to pinpoint precisely the different student performances which qualify for each Level of Achievement. Further, it is somewhat difficult to sort out exactly what are the criteria and what are the indicators of different performances relevant to each criterion. For example, is the criterion ‘accuracy in the application of definitions etc.’ or is it merely ‘application of definitions’? Similarly with ‘some ability to recall formulae and definitions’. Is this a criterion or is it a standard relating to the criterion of application of formulae and definitions?

It would appear that there are problems associated with packaging criteria and standards together in the same verbal description. By abstracting them and separating them out as detailed in the schematic examples of Method I, it is contented here that teachers may be able to judge with *a greater degree of precision* where particular student achievements fit.

Also the adherents of Method I usually incorporate combinations of standards and criteria or trade-offs and compensations in their schemas. This allows a degree of flexibility without jeopardising the comparability that must exist between the achievements of different students. By consulting the combination/trade-offs table in a subject, students should be able to see how different combinations of standards along the various criteria might be commensurate with their own abilities and interests. They have, in effect, a well-defined table to help them decide where they might most profitably expend their efforts in exchanging performances for an exit Level of Achievement. A Year 11 student of Music who is already an accomplished violinist and who is aiming for an exit level of VHA, could, on consulting the table in Step (iii) of the Appendix I schema, see that he/she needs at least one other top standard performance (apart from practical music-making), and two at ‘top-minus-1’ or standard 5. Depending on their strengths and weaknesses, students can aim at the combination of standards and criteria that is feasible for them and which will enable them to qualify for the Level of Achievement to which they aspire.

By way of summary, it is suggested here that the method of separating criteria and standards and incorporating a system of combinations or trade-offs is preferable to fusing criteria and standards together in verbal descriptions. In comparison with the latter, the former are likely to result in:

i. greater precision in documents that refer to criteria and standards;

ii. more precise judgements on where particular student performances fit on a well-defined ladder of achievement; and

iii. greater flexibility, more on-target expending of effort by student towards aspired-to Levels of Achievement and a clearer picture of allowable combinations of standards or trade-offs.

# Appendix1

***Table 1: music—exit level end year 12. Criterion identification [step (i)] and standards specification [step (ii)] (to be supplemented by the exemplars of students' work)**

LOWEST PROFICIENCY HIGHEST PROFICIENCY

| CRITERIA | STANDARD 1 | STANDARD 2 | STANDARD 3 | STANDARD 4 | STANDARD 5 | STANDARD 6 |
|---|---|---|---|---|---|---|
| CREATIVE WRITING SKILLS | Student has appropriated almost no musical styles and forms and shows little ability to make any sort of statement using these styles and forms | Student has appropriated few musical styles and forms and only occasionally can made an imaginative statement using these styles and forms | Student has appropriated some musical styles and forms and can sometimes make a reasonably imaginative though not always very personal statement using these styles and forms | Student has appropriated many musical styles and forms and can make a fairly imaginative and personal statement using these styles and forms | Student has appropriated a large number of musical styles and forms and can make an imaginative and personal statement using these styles and forms | Student has appropriated a maximum number of musical styles and forms and can make an original, imaginative and personal statement using these styles and forms |
| PRACTICAL MUSIC-MAKING | Student almost never displays a reasonable standard of musical performance and interpretation in either instrumental or vocal areas, and cannot sight-read, improvise of conduct. | Student rarely displays a good standard of musical performance and interpretation in either instrumental or vocal areas, can sight-read only with difficulty and conduct and improvises with reluctance and little competency. | Student sometimes displays a good standard of musical performance and interpretation in either instrumental or vocal areas and can sight-read, improvise and conduct when called upon but not always competently. | Student often displays a high standard of musical performance and interpretation in both instrumental and vocal areas (though perhaps not equally in both) and can also sight-read, improvise and conduct with a reasonable degree of competence. | Student displays a high standard of musical performance and interpretation in both vocal and instrumental areas and can also sight-read, improvise and conduct with competence. | Student displays exceptionally sensitive standards of musical performance and interpretation in both vocal and instrumental areas and can also sight-read, improvise and conduct with flair and skill. |

| CRITERIA | STANDARD 1 | STANDARD 2 | STANDARD 3 | STANDARD 4 | STANDARD 5 | STANDARD 6 |
|---|---|---|---|---|---|---|
| AURAL PERCEPTION | Student almost never displays any aural perception in any area and almost never completes any of the given tasks. | Student sometimes displays a fair level of aural perception but has trouble with both seen and unseen works; can only rarely perform given tasks with any ease. | Student displays a fair level of aural perception in seen work but sometimes has difficulty with unseen; can perform given tasks but not always without effort. | Student displays a good level of aural perception in seen and most unseen work and can perform most given tasks without too much effort. | Student displays a high level of aural perception in both seen and unseen work and can perform most given tasks with ease. | Student displays an outstanding level of aural perception in both seen and unseen work and can perform all given tasks with ease. |
| MUSICAL KNOWLEDGE | Student has an extremely limited grasp of musical knowledge in all areas (historical, theoretical and stylistic) and finds the greatest difficulty in discussing issues in any of these areas. | Student has a poor grasp of musical knowledge in all areas (historical, theoretical and stylistic) and has difficulty discussing issues in these areas with any degree of competence. | Student has a reasonable grasp of musical knowledge in some areas (historical, theoretical and stylistic) and can occasionally discuss issues in some of these areas with a degree of competence. | Student has a good grasp of musical knowledge in most of the areas (historical, theoretical and stylistic) and can discuss issues in some of these areas quite competently. | Student has an impressive grasp of musical knowledge in all areas (historical, theoretical and stylistic) and can discuss issues in most areas quite competently. | Student has an impressive grasp of musical knowledge in all areas (historical, theoretical and stylistic) and can discuss issues in these areas with insight. |

## *Table 2: trade-offs—policy on permissible combinations for the awarding of levels of achievement on exit [step (iii)]

(all four criteria are equally weighted in this example)

| LEVELS OF ACHIEVEMENT | COMBINATIONS OF STANDARDS TO BE SATISFIED |
|---|---|
| VHA | At this level, a student achieves at least *Standard 6 in two areas* and at least *Standard 5 in the other two areas.* |
| HA | At this level, a student achieves at least *Standard 5 in two areas* and at least *Standard 4 in the other two areas.* |
| SA | At this level, a student achieves at least *Standard 3 in three areas* and at least *Standard 2 in one area.* |
| LA | At this level, a student achieves at least *Standard 2 in three areas* and at least *Standard 1 in one area.* |
| VLA | At this level, a student achieves at least *Standard 2 in one area* and at least *Standard 1 in three areas.* |

*Source: Ms Lorna Collingrodge, MacGregor SHS, Visiting Teacher, Sem 1. 1985, Department of Education, University of Queensland

# Criteria and Standards in Senior Health and Physical Education

## Discussion Paper 14

**Abstract**: This paper is concerned with the assessment of students' global achievements in Senior Health and Physical Education. It examines the notion of global achievement in the subject and suggests the criteria and standards by which the quality of student performance can be judged. It also suggests specifications for awarding exit Levels of Achievement which reference the standards schema.

**Author**: Robert Bingham,. Assessment Unit, January 1987

***Note***: This Discussion paper has been produced by a member of the Assessment Unit of the Board of Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. It should not be taken as representing Board policy, and is being circulated within the teaching profession to obtain reactions to its contents.



## Introduction

Of primary concern under ROSBA are the criteria and standards to be used when assessing the quality of student performance. The program of assessment in each subject culminates in the judgment of global achievement using these criteria and standards and the subsequent award of an exit Level of Achievement. It is important therefore to set the specifications by which student achievement at exit from a subject will be assessed.

It is the intention of this paper to:

- define global achievement;
- produce subject-related criteria for Senior Health and Physical Education;
- nominate a set of standards on each criterion for global assessment; and
- establish specifications for the award of exit Levels of Achievement.

## Global Achievement

This is the summary which represents a composite of achievements over all the components of the subject (e.g. Theory and Practical components) that is in terms of performance over the whole course rather than in relation to separate components or units of work. The students' achievements are profiled on the common achievement dimensions which run through all the components of the subject and which are taken as characteristic of the subject. This process involves determining the standard attained on each criterion that most effectively represents the student's achievements in that dimension. Global achievement is then expressed as a profile of standards of performance attained on the subject-related criteria.

The comparison of this school-assessed profile of global achievements with the syllabus specifications for the award of exit Levels of Achievement determines each student's exit Level of Achievement. The foundational task for these global assessment processes is the derivation of a set of subject-related criteria and standards.

## Criteria and Standards Defined

As defined in other Assessment Unit discussion Papers (1–3, 1986), criteria are properties, dimensions or characteristics by which something is judged or appraised. These criteria are based on the characteristics of the subject or discipline and are implied by the course objectives. As subject-related criteria they should be continuously relevant throughout the entire course (Discussion Paper 3). They describe broad performance continua on which the range of student performances can be expressed from lowest to highest proficiency. The bench marks or points of reference along each criterion used to differentiate levels of performance are called standards. For example, one of the observable dimensions by which performance in the Practical component of Physical Education can be judged is ‘Skill Application’. Along this criterion, student performances could be located at different levels along the continuum. At one extreme are those performances which show little ability to perform the basic skills in simple drill situations; at the other end are those showing an ability to adapt difficult skill sequences to unusual occurrences in a game. A student who can perform basic combinations of skills with form, accuracy, and consistency performs in a qualitatively different manner from another student who can perform only individual skills with form, accuracy, and consistency. The two students would have therefore reached different standards on this skill criterion. Standards can be expressed as verbal descriptions of student performances on a criterion, as shown in Table 1.

### Table 1: Example of the standard criterion relationship

| CRITERION | STANDARD 1 | STANDARD 3 | STANDARD 5 |
|---|---|---|---|
| SKILL APPLICATION | Can perform the basic individual skills in drills with limited control at low speeds. | Performs most prerequisite skills and combinations of skills with form, speed, accuracy and consistency in complex drills. Attempts appropriate skills in mini-games where cues are distinct. | Performs the basic skills and combinations of skills with control in a wide range of ‘game’ situations. |

In practice, a number of criteria may be used simultaneously for assessing performance. Because there may be different numbers of standards distinguishable along each criterion the combination of criteria and standards can be expressed as a matrix or standards schema as in McMeniman (1986) and Findlay (1986).

## Deriving Criteria in Health & Physical Education

Work Programs utilise diverse objectives and criteria according to the different units of work. To date a wide diversity of criteria and a wide range of assessment tasks and instruments exist. These are traceable to a focus on the particular nature of the activities themselves in the course rather than on common features. This is particularly true in the physical activity units and it has precluded the development of a comprehensive, criteria-related concept of global achievement in the subject. It is not practicable to derive the criteria for global assessment by collapsing the existing unit-related criteria into a common set. It is important therefore to identify the common dimensions running through all components of the subject around which the course of study can be unified and which suggest the broad, subject-related criteria for global assessment.

As a subject Health and Physical Education has two major components:

a. Active participation in a range of physical activities (referred to in this paper under the heading of Games and Sports) which is commonly referred to as the practical component of the subject.

b. A theory component which encompasses the foundations of Physical Education and Health Education both of which draw on discipline-based knowledge and are considered primarily as academic pursuits.

The practical Component encompasses both the performance aspect and an essential experiential aspect. Students are exposed to a range of activities in such a way that they not only will become effective performers in a number of activities but also reflect on the experiences. Through analysis of the performance experience the student-performer identifies problems or areas of need in order to establish solutions for future occasions. Analysis and reflection are necessary in associating game theory with the experience and will influence strategy development and performance modification. This approach aims at producing the knowledge that allows considered decisions to be made about future participation, as well as producing the understanding and skills that will result in effective performance. Effective performers do more than reproduce the basic skills or operate on reflex action. They understand the total context of the activity and can adapt skills, strategies, and physical capacities to meet the demands of participation. Such performers know not only how to do things, and what to do, but also when, where and why to do them.

To be effective performers in the Practical component of the subject involves more than performing in this manner in isolated activities. It involves performing to a particular standard over a range of the activities included. Effective performers can identify the common features of, and differences between, various types of activities, and can generalise skills and strategies. This view acknowledges both the cognitive and psycho-motor components of effective performance and suggests the following dimensions as salient to performance in Games and Sports and as criteria by which achievement in the practical component of the subject may be judged:

- Skill application;
- Knowledge about Games and Sports (where Games and Sports are considered in the widest sense not just of as those included in this Element in the present syllabus);
- Employment of Strategy

Some of these encompass a number of sub-dimensions which are interrelated and contribute to the broader dimension (see Figure 1).

### Figure 1. Contribution to Achievement in the Practical component

ACHIEVEMENT IN THE PRACTICAL COMPONENT

| KNOWLEDGE ABOUT GAMES AND SPORTS | EMPLOYMENT OF STRATEGY | SKILL APPLICATION |
|---|---|---|
| Recognition of relations, roles<br>Recall of rules, strategies, etc.<br>Representation of positions, game occurrences etc.<br>Explain strategies. | Positioning<br>static<br>dynamic<br>Use of space<br>Distance<br>Angle of Attack<br>Direction of Movement<br>Interaction with others<br>Team Play<br>As opposition | Skill Production and Movement Control<br>Form/Style/Technique<br>Accuracy<br>Speed<br>Consistency<br>Adaptability<br>Skill Selection<br>Appropriateness<br>Creativity |

For example, Skill Application encompasses sub-dimensions of Skill Production and Skill Selection. Skill Production considers the voluntary control over the movements needed to reproduce a particular action and is evaluated according to the accuracy of the movement, the correctness of form or technique, the speed of production, and the ability to adapt skills to changing requirements. An overriding consideration is the context in which the student can perform these skilled movements. The situations under which the skills need to be produced range from simple drills (which require responses to simple and unambiguous cues and are performed in isolation from the applied context) to the complex ‘in-game’ situations (with large displays of potentially viable cues and severe time constraints). Skill selection involves the interpretation of these contextual cues and a decision as to which skill or skills to use. Effective skill selection is revealed in the appropriateness of the skills attempted to the context.

Similarly, the Employment of Strategy is composed of the sub-dimensions of Positioning, the Use of Space, and the Interaction with Others. While the Employment of Strategy applies to all activities, some of the sub-dimensions may not. Just as the concept of Strategy differs with different classes of activity, so the connotations attributed to the sub-dimensions may need re-interpretation according to the class of activity. In an activity such as Rugby, the Use of Space may be interpreted differently from its interpretation in an activity such as Gymnastics. Similarly, the Interaction with Others is only marginally relevant in Archery, whereas in Rugby it would include both team play and the reaction to opponents.

Knowledge about Games and Sports involves more than knowledge of the rules as reflected in students' responses to the decisions of officials. It involves the knowledge of game structures, role requirement and strategies which produce an intelligent, effective performance. Assessment of practical performance therefore involves imputing intent and understanding to the actions observed. However it may not be sufficient to take responses to game occurrences as the sole evidence of Knowledge about Games and Sports. Practical situations are unlikely to provide the range of events and the changes in positional roles for revealing the full extent of a student's knowledge. It is not sufficient to take performance alone as evidence of Knowledge about Games and Sports. It may be also important to use written assessment, particularly for students whose physical performance belies their cognitive understanding.

The Theory component encompasses both the Foundations of Physical Education and Health Education. The Foundations component is drawn from the sub-disciplines that address

- the structures and functions of the human body;
- sports medicine; and
- the social and historical contexts of games and sports.

This component aims at producing students who understand concepts and principles from these sub-disciplines and who know how, why, and when to apply them. Students will be able not only to recall information but also to solve movement-related problems. As the focus on solving practical problems is increased, the range of knowledge and academic skills required to provide and explain the solution increases. Students therefore have not only to research certain movement events but also to report their conclusions in ways appropriate to the sub-discipline. Problem solving also requires the integration of the contributions from the sub-disciplines in an interdisciplinary approach appropriate to Physical Education.

Similarly in Health Education there is discipline-based knowledge from Anatomy, Physiology, Health, Social and Preventive Medicine that is used to solve problems associated with health conditions and lifestyle.

- It is proposed that achievement in the theory component be based on the following criteria:
- Knowledge about the Foundations of Physical Education and Health Education Application of Knowledge in Problem Solving
- Academic and Research skills.

These broad dimensions of achievement can also be divided into sub-dimensions (see Figure 2) which contribute to achievement in the component and may have particular importance in different task contexts.

### Figure 2. Contribution to Achievement in the theory Component

ACHIEVEMENT IN THE THEORY COMPONENT

| KNOWLEDGE ABOUT FOUNDATIONS OF P.E. AND HEALTH EDUCATION | APPLICATION OF KNOWLEDGE IN PROBLEM SOLVING | ACADEMIC AND RESEARCH SKILLS |
|---|---|---|
| Recognition of information.<br>Recall definitions.<br>procedural constraints.<br>Comprehension:<br>Explain relationships<br>Compare and<br>.Constrast principles, structures. etc.<br>Analysis and Synthesis of Information. | Identification of difficult problems<br>Identification of features relationships, patterns. etc.<br>Selection of potential solutions and adaptations.<br>Derivation of criteria for judging solutions. | Accuracy.<br>Appropriateness of conventions.<br>Presentation/Visual Appeal.<br>Selections and application of appropriate procedures and techniques. |

## Deriving the Global Standards Schema

A standards schema for Health and Physical Education is intended to function as a basis for expressing students' achievements in the subject as a whole in terms of common performance criteria. It should enable schools to provide, from the criteria-related data collected, a profile of each student's global achievements that is comparable with other student profiles in the subject across the state. The schema therefore has to be explicit enough for schools to tailor their assessment programs in such a way as to collect data for determining a student's location on the global schema.

The criteria by themselves are simply dimensions of performance and are not sufficient for differentiating levels of student achievement. The specification of standards on each criterion allows the performance of students to be differentiated on each dimension. The standards therefore supply explicit descriptions of performance specifications and tap the sub-dimensions of which the criteria are composed. The upper and lower standards on a criterion denote the limits of the performance continuum operational at the level of schooling. But this schema is for assessment of global achievements in the subject presumably performed at the end of a two-year course. While the performance criteria may be continuously relevant throughout the entire course, the standards are not. Although the standards on a criterion can be interpreted as part of a performance continuum this schema does not present a developmental sequence over the course. The standards focus on the quality of achievements expected after two years study in the subject and are expressed in the context of the total course. The combination of standards over all the criteria form a matrix or standards schema as presented in Appendix 1.

The school is responsible for determining the standard attained on each criterion that most effectively represents the student's achievements in that dimension. The notion of a 'best-fitting' standard for each criterion suggests a number of approaches to summarizing the assessment data, including a 'typical performance' approach which focuses on standards generally evidenced on the dimensions across tasks and course segments. In the Practical component this means performing to this standard generally across the majority of activities involved. The student profiles in the Practical and Theory Components are established independently but together comprise the global profile provided by the school.

Although the summary mechanism is not the major interest of this paper it should be noted that this concept of global achievement focuses on criteria and standards and does not consider the repeated compression or the final arithmetic summation of date.

The contextual relationship evidenced in the standards encompass both a concept of the course as a whole and the actual performance context. The schema assumes a unified subject in which subject-structure, core concepts, teaching strategies and learning experiences promote the common threads which run through the major components of the subject. Therefore an overarching requirement that surfaces through the standards and the description of global achievement as a distinguishing feature of high standard students is that they evidence understanding of the unifying links and perform at the standard level in more than isolated fragments of the course. Their achievements not only evidence consistency across a range of tasks but show an ability to interrelate units of work and task features.

The standards expressed on the criteria for the Practical component are intended to take into account the interaction between student abilities and the performance context. Embedded in these standards are expressions of the complexity of the situations in which students may show a particular quality of performance. The standards expressed on the Theory criteria describe increasingly complex student achievements in relation to knowledge and methodologies drawn from the sub-disciplines.

While it is important to make the schema specifications broad enough to accommodate a range of evidence it is also clear that they have to be specific enough to enable comparability of between school judgments of global achievement despite some variations in the specific learning experiences and assessment tasks. The relevance of the standard schema will be enhanced with the production of exemplars.

Detailing standards on the global criteria in a standards schema, as in Appendix 1, should not only allow the global assessment to be made but will assist those interim assessments during the course. The standards need little explications to make them specific to the unit level assessments. For the Practical component the standards need to be explicit, because they are intended as the framework by which teachers observe and categorise student performance in the applied contexts of the activity. The assessment schema for these interim assessments will have to be internalised in order to make 'in-situation' judgments of the quality of performance.

## Determining Exit Levels of Achievement

Student achievements in a Board subject are ultimately certified by the BSSS and expressed in a single, summary statement as an exit Level of Achievement. To achieve state-wide comparability and facilitate certification, there must be a common mechanism by which students' global achievements in a subject are awarded an exit Level of Achievement. The subject syllabus should specify, as profiles of global achievement in the subject across the standards-schema, the particular performance combinations appropriate to each of the five exit Levels of Achievement. The global profile provided by the school can then be compared with these syllabus specifications for the award of the exit Level of Achievement. It seems inappropriate to set a single fixed prescription for each Level of Achievement as a number of students' performances may be judged equivalent although they are expressed as different combinations of standards on the criteria. One way of accommodating these equivalent-but-different global achievements within the comparison mechanisms is to establish the profile of standards on the schema which must be 'at-least' met for each exit Level of Achievement. The 'cut-off' for each Level is therefore expressed by standards-on-criteria.

### Table 1: A table for awarding exit Levels of Achievement.

| LEVEL OF ACHIEVEMENT | VLA | LA | SA | HA | VHA |
|---|---|---|---|---|---|
| CRITERIA | | | | | |
| Skill application | | S2 | S4 | S5 | S6 |
| Employment of strategy | | S2 | S2 | S3 | S4 |
| Knowledge about games and sports | | S2 | S3 | S4 | S5 |
| Knowledge of foundations | | S1 | S2 | S3 | S4 |
| Application of knowledge in problem solving | | S1 | S2 | S3 | S4 |
| Academic and research skills | | S1 | S2 | S3 | S4 |

The table gives the minimum performance composite expected for each exit Level of Achievement and indicates a performance range between achievement levels. VLA students obviously have no minimum profile but achieve below the minimum set for LA. Criteria-based assessment and use of the standards schema still allows comment on VLA students' actual performances both for specific units of work and the course as a whole. A student's global achievement profile can be compared with this table and the exit level determined.

**For example:**

A student with a profile of S3, S2, S2, S3, S2, S2 across the standards schema would be awarded Limited Achievement as the student profile does not meet the 'minimum requirement' profile of standards for SA. Although it meets the SA requirements on some criteria and exceeds them for one criterion it does not meet the standard requirements for SA on 2 of the 6 criteria.

This 'minimum profile' or 'at-leasts' model does not allow for a trade-off in which outstanding performance on some criteria may balance out weaker performance on less important criteria. If such a trade-off notion were to be considered the syllabus would have to state the emphasis placed on some criteria relative to others and the resultant trade-offs allowed.

## References

Findlay, J *Mathematics Criteria for Awarding Exit Levels of Achievement*. Discussion Paper 7, B.S.S.S., 1986

McMeniman, M. *A Standards Schema*. Discussion Paper 3, B.S.S.S., 1986.

Sadler, R. *ROSBA's Family Connections*. Discussion Paper 1, B.S.S.S., 1986. The Case for Explicitly Stated Standards. Discussion Paper 2, B.S.S.S., 1986.

## Appendix 1: The Standards Schema for Senior Health & Physical Education

The first half of the schema expresses expected global performance in the Practical component over the three criteria:

- Knowledge About Games and Sports;
- Skill Application; and
- Employment of Strategy

After profiling student achievement on the Practical component according to this schema, then the second half of the schema which expresses expected achievement in terms of:

- Knowledge About the Foundations of PE & Health Education;
- Application of Knowledge in Problem Solving; and
- Academic and Research Skills;

can be employed to determine global achievements in the theory component of HPE.

## Schema for practical component

| *CRITERION* | *STANDARD 1* | *STANDARD 2* | *STANDARD 3* | *STANDARD 4* | *STANDARD 5* | *STANDARD 6* |
|---|---|---|---|---|---|---|
| KNOWLEDGE ABOUT GAMES AND SPORTS | Recalls the basic rules, positions and ethics. | Recalls and applies the basic rules and ethics in common situations in games and sports | Recalls and applies a wider range of rules and ethical procedures to appropriate situation.<br>Recognizes player positions and common pattern of play or performance associated with games and sports. | As for Std 3 plus: Discusses relationship of patterns of play, strategies, and player-positions to specific classes of games and sports. | As for previous standards plus: Compares and contrasts features of games and sports across the classes of games and sports. | |
| SKILL APPLICATION | Can perform the basic individual skills in simple drills with limited control at low speeds | Can perform basic individual skills and simple combinations with form and accuracy at slow speeds in drills. | Performs most prerequisite skills and combinations of skills with form, accuracy, speed and consistency in complex drills. Attempts appropriate skills in minimum games where cues are distinct. | Perform the appropriate skills with control in mini-games (e.g. 3 on 3,4, and 4) or similar controlled situations. Attempts appropriate skills in common 'game' situations. | Performs the appropriate skills and combinations of skills with control in a wide range of 'game' situations. | As for S5 over a wide range of activities. Adapts skills to changing requirements |
| EMPLOYMENT OF STRATEGY | Positions self at start of play/activity and restarts. Coordinates with others in simple drills. | As for S1 plus: Positions self in set positions in set plays. Coordinates with others in well rehearsed plays. | Shows basic understanding of positioning in relation to field space, team mates and opponents in common situations. | Across a wide range of situations the student shows a strategic awareness of space and positioning. Anticipates plays and supports team mates in offence and defence. | Performs as for S4 across the range of games and sports. Anticipates plays and initiates team responses. | |

## Schema for theory component

| *CRITERION* | *STANDARD 1* | *STANDARD 2* | *STANDARD 3* | *STANDARD 4* | *STANDARD 5* |
|---|---|---|---|---|---|
| KNOWLEDGE ABOUT FOUNDATIONS OF P.E. AND HEALTH EDUCATION | Recalls basic facts (definitions and principles) from the major areas of study. | As for S1 plus: Recognizes important features of cause-effect relationships and recalls pertinent facts. | As for S2 plus: Recalls theoretical concepts and explains basic cause–effect relationships across areas of study. | As for S3 plus: Shows understanding of a wide range of concepts. Can discus the relationships between discipline-based explanations phenomena. | |
| APPLICATION OF KNOWLEDGE IN PROBLEM SOLVING | Needs extensive teacher assistance to identify features of problem and to relate procedures to situation. | Can follow teacher direction to establish features of problem and to institute appropriate solution. | Independently identifies key features of cause–effect relationship in problems and provides procedural solution. | Independently identifies the key features of a range of problems and establishes criteria for acceptable solutions. | Relates theoretical principles to key features of practical problems across all area of study and provides solutions to problems. |
| ACADEMIC AND RESEARCH SKILLS | Gathers 'data' under teacher supervision and records and reports on established forms. | Can follow documented procedures and conventions for data collection, interpretation, and reporting. | Independently uses standard research procedures and reporting modes and interprets normal range data. | Shows understanding of a range of research procedures and reporting modes and interprets a range of data across the areas of study. | |

# Improving the Quality of Student Performance through Assessment

## Discussion Paper 15

**Abstract**: Two basic assessment mechanisms through which the quality of student performances can be improved are feedback and information supplied about task expectations prior to performance. In this paper the complementary nature of these two mechanisms is examined, while feedback is analysed to indicate the value of certain forms of feedback over others. The paper complements and further develops some ideas concerned with formative and summative assessment presented in an earlier Discussion Paper.

**Author**: Janice Findlay, Assessment Unit, January 1987.

***Note:*** This Discussion Paper has been produced by a member of the Assessment Unit of the Board of Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. It should not be taken as representing Board policy, and is being circulated within the teaching profession to obtain reactions to its contents.



## Introduction

Many factors, both internal and external to the school, affect the quality of student learning, including needs, interests, abilities and motivations of students, family situations, and teacher attitudes and approaches. Within the curriculum, assessment can be a highly effective, albeit not the only, mechanism used to promote improved student learning and performance.

When used formatively, assessment is an important technique through which teachers seek to improve student learning and subsequent achievement. In addition to the assessment instrument itself, the provision of task expectations in terms of assessment criteria and standards, prior to the performance of the task is supportive of improving student achievement. Further, exemplars of what constitute varying standards of quality for individual tasks, and of acceptable combinations of standards on criteria over a variety of tasks for a particular Level of Achievement, also can supplement prior expectations and actual performance feedback.
The first two aspects of assessment form the major focus on this paper.

## Formative and Summative Assessment

The terms ‘formative assessment’ and ‘summative assessment’ have evolved from the terms ‘formative evaluation’ and ‘summative evaluation’ which were first applied to curriculum evaluation

by Scriven (1967). Formative and summative assessment, however, apply to student assessment. McMeniman (1986) defines formative and summative assessment respectively, in the following way:

Formative assessment occurs when assessment, whether formal (e.g. testing) or informal (e.g. classroom questioning), is (a) primarily intended for, and (b) instrumental in, helping a student attain a higher level of performance.

Summative assessment is designed to indicate the achievement status or level of a student. It is geared mostly towards reporting at the end of a course of study, especially for purposes of certification.

## Improving Student Performance Through Assessment

Two complementary assessment mechanisms which assist in the improvement of student performance are feedback and prior specification of task expectations.

## Feedback

Feedback is the principal mechanism through which assessment for formative purposes is realised. It refers to information returned to the student AFTER tasks have been performed, and is idiosyncratic, that is, feedback is referenced to the specific performance of a student. Feedback is not only an integral aspect of formative assessment but an essential element of the teaching-learning process. Further, it involves more than information, printed or otherwise, concerning the correctness or incorrectness of a response. To be effective, feedback requires (among other things)

(a) information about the position of the student's performance in relation to the specified criteria, stated in the prior information provided to students (the reference goal); and

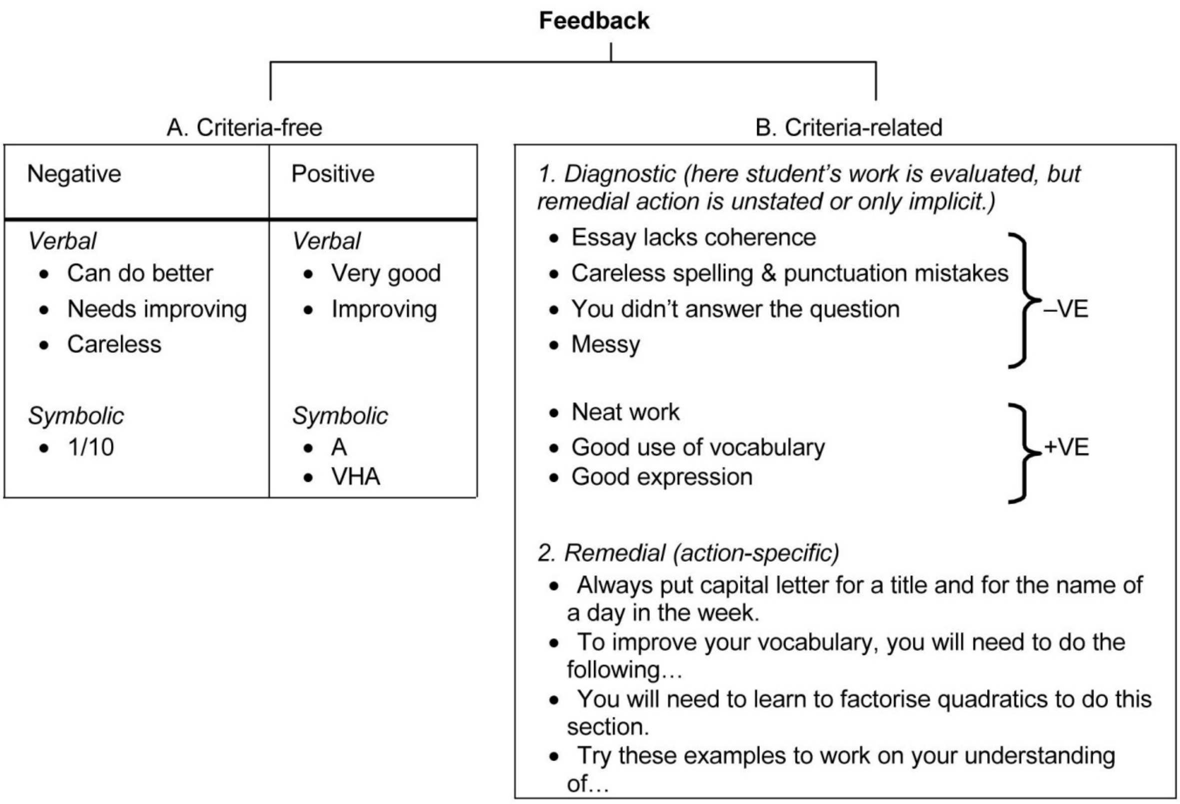
(b) suggested appropriate actions to decrease the gap between the two positions.

Quality, frequency, and amount of feedback can influence the rate of learning, and ultimately, the Level of Achievement of a student. Any improvement in these aspects of feedback can be seen as affecting positively the chances of improving student achievement.
It is useful to classify various forms of feedback in terms of their relationship to criteria-based assessment as shown in Figure 1. The major dichotomy is between whether or not feedback is referenced to specified criteria. Within the classification that is 'criteria related', a further significant delineation is between feedback which provides information on present status only, and feedback which also presents information about remedial action/s required to improve performance. Such classification has clear implications for decisions about the form of feedback provided to students.

The *sole* use of 'criteria-free' feedback lacks specific dimensions to which a student can refer, to assess learning. Further, such feedback does not attempt to direct a student's efforts towards improvement. That is, remedial action is not necessarily forthcoming.

Desirable feedback contains both diagnostic and remedial criteria-related information. It includes identification of the problem areas, suggested ways of solving or decreasing the problems, and some comments on positive features of the performance. Educationally, this last type of information is essential to continued learning because its 'praise' dimension can be directly related to motivation, and the experience of success. Other combinations such as positive criteria-free feedback together with the remedial therefore, may also prove effective, particularly when there is emphasis on remedial actions aimed at improving the learning.

**Figure 1: Classification of Feedback**




All the feedback categories listed in Figure 1 occur to varying degrees within existing practice. However, for improved feedback, more criteria-related feedback containing information about remedial action as well as diagnosis should be conveyed to students.

## Prior Expectations of Performance

Various methods can be used to provide information to students about both immediate standards to be attained, and exit standards. During the public examination era, past papers provided an important avenue through which students were exposed to the examiner's expectations and could develop strategies for decreasing the gap between their own level of achievement and the one aspired to. Past papers however, never provided any information as to how the gap could be lessened. Annotated exemplars of various standards of work are another way of providing students with information about reference points, and the qualities these exhibit. In all cases however, teacher input is crucial in determining what students need to do in order to achieve at higher levels, because this information is idiosyncratic.

One major advantage of the proposed criteria-based system is the requirement to prespecify and make publicly available the criteria and standards to be used to judge student achievement at exit. This applies at a more basic level as well. Information provided to students *prior* to the performance of tasks (sometimes called feedforward) is complementary and essential to useful feedback, that is, criteria-related feedback. Conditions under which tasks are to be performed such as length of task, weighting, content areas to be sampled, and marks allocated to individual questions, generally have always been an integral part of pre-task information. However in a criteria-based context, this prior information not only includes specifications about the task, and about the conditions under which it is to be performed, but also criteria to be used for assessment. It is not idiosyncratic as is feedback, but has a 'broadcast' flavour arising from its generality and availability to all students.

Figure 2 shows an example of such information accompanying tasks. It is a modification of an example provided by Grech (1986).

**Figure 2: Task conditions and criteria**

| **Media Response: 'The Sting'** |
|---|
| *Task*: In an attempt to discover the greatest American film of all times, the 'Los Angeles Voice' is holding a competition in which readers are invited to submit critical reviews of appropriate films. You have decided to enter a critique of 'The Sting', hoping that your submission will be published as the winning entry. You have decided to highlight the following features of the film:<br>(a) the intricacies of plot structure<br>(b) social comments made about the era<br>(c) characterisation and acting ability<br>(d) special effects and realism<br>(e) filming technique<br>(f) overall effectiveness and entertainment values<br><br>Length: 500 words<br><br>Weighting: 10%<br><br>Conditions: Students will view the film in class time, start the assignment at school, and finish it at home |

| **Marking criteria** | |
|---|---|
| *Criteria* | *Comments* |
| 1. understanding & completion of the task<br>2. ability to view critically<br>3. ability to express ideas in an organised manner<br>4. ability to form & justify opinions<br>5. maturity & originality of ideas<br>6. technical proficiency<br>7. understanding of the film<br>8. presentation | |
| | |
| Overall comment: | |

The example is taken from the subject area of English, which along with some other disciplines such as Foreign Languages, has been using successfully such an approach for years. In these cases the implementation of ROSBA has often led merely to a formalisation of previous practices. For some subject areas however, specification and inclusion of definite criteria for assessment, as distinct from assessment conditions will be a new dimension to be evolved.

Criteria and standards for specific task assessment provide students with short term, tangible targets. Global criteria and standards, (and accompanying annotated exemplars), used for awarding exit Levels of Achievement are long term targets. Both short and long term targets allow students to pursue degrees of excellence with more awareness and efficiency. In addition, such prior information introduces a visible form of justice, in the sense that no student is left uninformed as to the expectations set by the teacher. That is, all students are 'equal' in terms of their access to prior knowledge about the task/s.

When criteria are provided prior to the performance of the task, there are some positive consequences. Repeated statement of these for each student is no longer necessary. The time spent devising these 'universal' task expectations and publishing them for all students should become less than the total time spent in providing these expectations repeatedly as part of the feedback process for each student. During feedback then, only information about the student's achievement and possible remedial actions need to be supplied for each student. Further, with task expectations specified, it may be possible to report much of the remedial feedback verbally to students in whole or small groups, or individually, and request that they write down the advice supplied. The amount of recording for official records may not be necessary if student folios exist and are well maintained. In the long term, the *total* amount of time spent on providing information to students both before and after performance of tasks, may decrease if prior information and these other strategies can be employed.

By incorporating both explicit statements of criteria for assessment and feedback, both the efficacy and efficiency of improving student performances are increased. This combination is preferable therefore to a system which utilises only one of these two complementary operations. In deciding the form and extent of prior information provided to students, schools naturally will consider the amount and types which provide optimum value to students while not depleting teachers' reserves of energy and time.

## Conclusion

A major goal of education is to improve students' learning. Feedback is one mode through which this is attempted. Feedback in the assessment context can exist in various forms. A preferred form has the following dimensions: reference to criteria, some favourable comments, and information on actions which effect improvements. This often is time-consuming. The complement of feedback, the prior specification of criteria and standards, seeks to reduce time spent on the first dimension of feedback and to allow both student and teacher to concentrate more on improving the former's achievement level. By incorporating both feedback and prior information, the quality of student performance should improve.

## References


Grech, D., Personal communication, 1986.

McMeniman, M., 'Formative and Summative Assessment — A Complementary Approach', Board of Secondary School Studies, January 1986.

Scriven, M., '*The Methodology of Evaluation*'. In Ralph Tyler et al. (eds.), Perspectives of Curriculum Evaluation, AERA Monograph Series on Curriculum Evaluation, No. 1, Chicago, Rand McNally, 39–83, 1967.

Sadler, D. R., 'Evaluation and the Improvement of Academic Learning', *Journal of Higher Education*, 54, 60–79, 1983.

Sadler, D. R., 'Theory of Formative Assessment' (Department of Education, University of Queensland, unpublished manuscript), 1983.

# A Pathway of Teacher Judgments: From Syllabus to Level of Achievement

## Discussion Paper 16

**Abstract**: This paper traces the decision-making process of teachers which allows information about student achievement within a course of study to be profiled over time. It attempts to place in perspective the different levels of decision-making and identifies the accountability of such judgments within the accreditation and certification process.

**Author**: Warren Beasley, Assessment Unit, January 1987

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Board of Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. It should not be taken as representing Board policy, and is being circulated within the teaching profession to obtain reactions to its contents.



## Introduction

When a school decides to offer a Board Subject, it takes steps to structure a course of study which satisfies the Syllabus requirements. The Syllabus is therefore the starting point for a series of judgements about the conduct of that course of study. These judgements culminate at exit from this course of study in the award of a Level of Achievement for each student. The broad guidelines promulgated in the Syllabus provide the boundaries within which the judgment about the Level of Achievement is to be made. Thus, the series of judgments which teachers make about a course of study also finish with the Syllabus. In other words, the syllabus, being an official statement from the Board of Senior Secondary School Studies, provides the authoritative framework for teacher judgments over time. This paper discusses the nature of those judgmental processes.

## The Structural Framework

The Board of Secondary School Studies is the statutory authority responsible for:

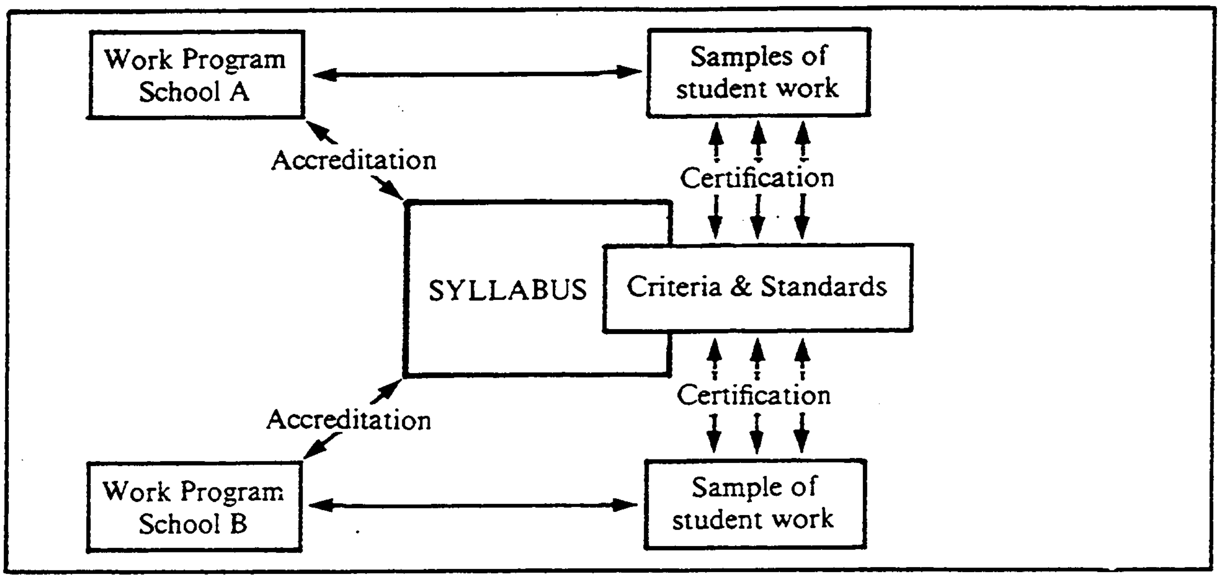
i the accreditation of school courses of study based on Board Syllabuses (Board Subjects) and courses of study proposed by a school on its own initiative (Board-Registered school subjects); and

ii the certification of student achievement in Board Subjects made by schools using review procedures laid down by the Board.

Within the context of the criteria-based philosophy of ROSBA (see Discussion Papers 1-4), the proposed framework for school and teacher decision making is outlined in Figure 1.

From a Board perspective, this framework provides a structure for both accountability and comparability. The Review Panel mechanism (see Handbook of Procedures for Accreditation and Certification, 1985) is concerned with those teacher judgments checked through the process of accreditation and certification. For students however, ROSBA operates in a broader educational context. Schools set out to provide a variety of learning experiences inside and outside conventional classrooms leading to much wider goals valued by the various educational systems and society generally. The school provides for the balanced social, physical and emotional development of all students. It is within this total educational context that the judgments of teachers are firmly located. The judgments which provide for the broad educational benefits accruing to students are represented mainly by the space outside the more formal boundaries of the procedures of Board accreditation and certification. This paper is concerned essentially with the decision-making which enacts ROSBA in the classroom. However, it recognizes that teachers see students in a context which is much broader than that represented in an accredited formal assessment scheme and that teachers draw upon their professional knowledge to inform, construct and implement instruments for the formal assessment program. This is experiential knowledge collected and assimilated in the day to day operations of classrooms. It includes in an informal way a knowledge of the students' performance characteristics as these evolve throughout the course of the study.

**Figure 1: Diagrammatic Representation of the proposed ROSBA accreditation and certification procedures (Double arrows signify comparisons that are made)**




## Teacher Judgments in the Educational Context

In an attempt to portray teacher judgments diagrammatically, a segment within a course of study is taken as the unit of analysis. This unit does not necessarily represent a semester of schooling. An interpretation of the relationship between the types and levels of teacher judgments within this unit is represented in Figure 2. The teacher judgments are reference to levels of criteria developed in Discussion Paper 9.

The relationship of teacher judgments within the context of the segment of work and the context of exit assessment is represented in Figure 3. Information which is referenced to Level 2 criteria is profiled over time from each segment of the course. At the end of the course, these information summaries are interpreted against global criteria and standards.

**Figure 2: From learning experiences to Levels of Achievement: a representation of teacher judgments over time.**

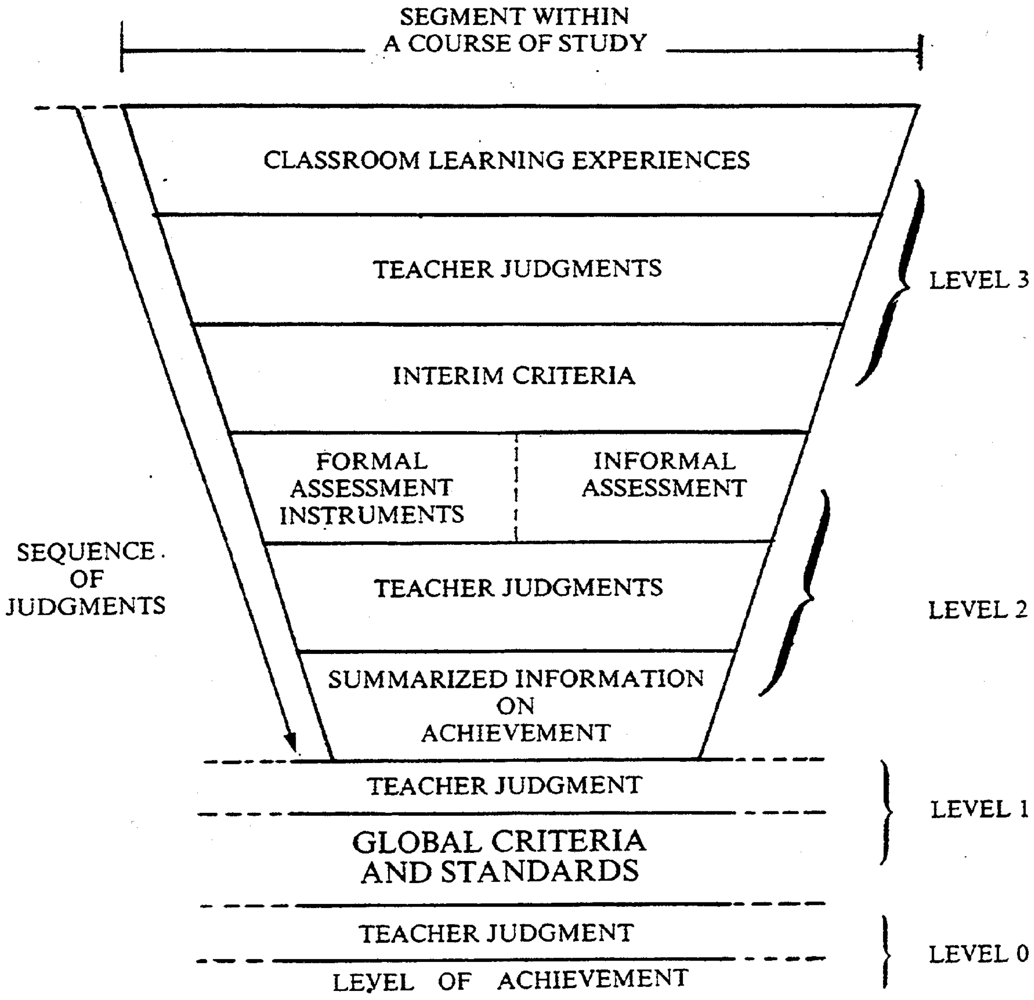
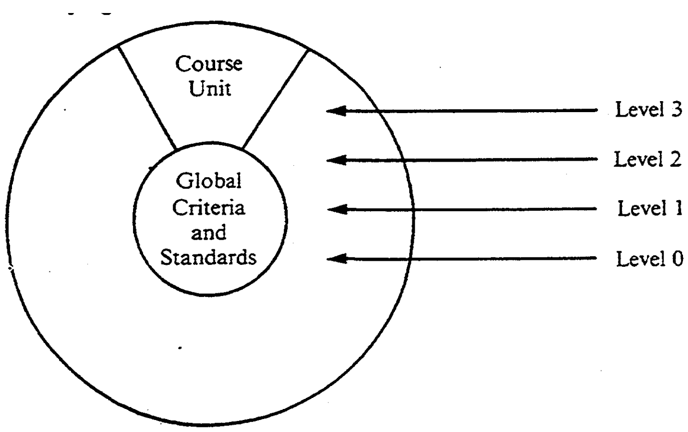



**Figure 3: Relationship of teacher judgments within a segment of a course and at exit.**



## Course Work, Assessment Instruments, and Teacher Judgments — Level 3

For meaningful learning to occur, students construct their own thought patterns and knowledge structures from what is taught, in the light of their own existing cognitive frameworks. This makes learning an individual affair which requires different learning styles, amounts of time, and effort, for each learner. For knowledge of learning outcomes to be beneficial to the student in the ongoing learning process, account must be taken during assessment of individual performance and communicated to the student. The judgment of the teacher is crucial in this process. Teachers determine the extent of correspondence between intentions in Work Programs (objectives, learning experiences, resources, etc.) and the learning outcomes demonstrated by students. Although Work Programs do attempt to prescribe in advance the outcomes of instruction, the process by which students 'learn' mediate such an 'objectivist' view of the world. The first hand experience of teachers in the day to day observation of student performance contributes substantially to any reliable projection of learning outcomes. The prescriptions of a Work Program cannot guarantee that the curriculum as received by students actually matches the planned curriculum. This fine-grained knowledge of students possessed by teachers should be used in the construction of the assessment instruments. At the level of individual classrooms, teacher judgments provide the most comprehensive measures of learning outcomes. The specific objectives judged by schools to be representative of the syllabus intentions and written into Work Programs represent the criteria (Level 3, Figure 2) through which the curriculum is enacted in the classroom. For the teacher in the day to day conduct of lessons, the collective importance of these Level 3 criteria manifests itself in the structured learning settings. By implication, these classroom activities lead to a set of outcomes, many of which have their source in an accredited Work Program.

On the other hand, these same criteria typically are much less explicit in the minds of students, because students construct knowledge in their own heads, building it into, extending, and modifying, their previous knowledge structures. This results in patterns of learning which do not necessarily correspond exactly with the pattern planned in the Work Program or as enacted in the classroom. It is the interaction of teachers with students which provides the teacher with insight into the learning outcomes, be they congruent or otherwise with the Work Program.

This intimate knowledge which teachers have of their students should form part of the information base used to determine Levels of Achievement. Formal assessment procedures adopted by schools should attempt to consolidate what teachers know about students and what students know about themselves. Assessment procedures should tell what students are capable of doing, and the design of instruments should allow this rather than measure what students do not know. Failure on a test instrument is not conducive to positive attitudes towards self or to an achievement oriented innovation like ROSBA. Assessment procedures, informal and formal, must contribute to the improvement of teaching and learning and should serve to ensure that the majority of students learn what the school system regards as the important tasks and goals of the education process.

## Assessment — An Exercise in Generalisation

The design of any assessment instrument which serves the assessment program for summative purposes (see Discussion Paper 6) is influenced by two requirements. On the one hand, teachers are attempting to provide a variety of avenues for students to demonstrate what they are capable of doing within the immediate context of the course structure. On the other hand, the certification process generally focuses on the general objectives located in syllabus documents. These parallel requirements are manifested in the choice of items in the assessment instruments and the interim criteria which teachers select to assess the student responses to that instrument (Level 2, Figure 2).

What is assessed formally can represent only a small proportion of the desired learning outcomes, but it is intended to reflect the focus of the teaching/learning settings provided in the course and to inform the teacher judgments about levels of student performances. This formal assessment procedure allows for the evolution of useful information about student achievement which is intended to be carried forward. Further it is used in conjunction with other information, gathered formally and informally from other course segments, and accumulated over the time of the course of study.

All the information from formal tests and teacher interaction collectively merges into a condensed form to become an interim report of the performance characteristics of students. This summary, preferably qualitative but often quantitative, generalizes the attainment of the course work objectives at Level 3 at some time prior to exit from the course of study. These teacher judgments using Level 2 criteria imply two types of performance characteristics — immediate and end of course: immediate in the sense that performance characteristics relative to short term interim standards are provided; long term in that the performance summary will be carried forward and used in conjunction with other information generated over the period of the course. The interim standards of student performance used depend upon the nature of the subject, its structural representation in a syllabus document, and the sequence in which the content of the course is presented to the students.

## Assessment of the Early Exit Student

Course structures have been described as being either ‘unitized’, for example, Senior Mathematics and Geography, or ‘developmental’ as in the case of English and Typewriting. Some could argue that the ‘content’ in unitized subjects is unique to each semester and each semester contributes equally to the measure of the knowledge of course content. But, given that intellectual growth is a developmental process, the argument that higher order cognitive thinking (referred to as ‘process’ in most syllabuses) remains isolated and is not developmental, is difficult to sustain.

For students who exit from a course before the end of Year 12, the interim standards become the basis for the award of an exit Level of Achievement. To use those standards (for example, at the end of Semester 1) for determining an exit Level of Achievement when learning outcomes are developmental in nature requires a high-order inference.

For unitized course structures, teacher judgment about an ‘early exit’ Achievement Level should reflect the corresponding subset of global criteria, appropriate for that unit and the associated standards. However very early exit from a developmental course requires a judgment reflecting more closely the interim criteria and standards. It is normally inferred in the latter case that student intellectual development will be maintained at a predicable rate and the exit Level of Achievement award can be extrapolated from the performance of students over a limited time frame of one, two or three semesters.

## Towards an Exit Level of Achievement

For the majority of students who complete the course as described in a syllabus document, the Level 1 criteria are the ones which focus teacher judgments towards the appropriate award of an exit Level of Achievement. This level of teacher judgment requires consensus from a more representative group than from within the school itself. Further, these Level 1 criteria reflect a ‘holistic’ view of the intended learning outcomes after experiencing the course of study and would normally be described in syllabus documents. These statements in the syllabus are intended to facilitate the maintenance of standards and an appropriate level of comparability across the State. Additional details concerning comparability within the context of ROSBA are provided in the ‘Discussion Paper 12, Defining and Achieving Comparability of Assessment’ (1986).

Teacher judgment at this level reflects Board policy on Exit Assessment (Fullest and Latest Information, Selective Updating, etc.) takes account of a profile of and any trend in the performance characteristics of students over the course of study, and matches this profile against the mandatory syllabus criteria and standards to be used for awarding exit Levels of Achievement. This final judgment is unique in that it shows to what extent students in a particular school have achieved standards of performance which are congruent with stipulated standards to be applied state wide. Although each pathway leading up to a

Level 1 judgment is unique to the school, inevitably and logically the starting point and finishing point of this judgmental pathway always resides in the syllabus.

## Conclusion

The discussion above highlights the importance of teacher judgments within a system of school-based assessment using the notion of criteria and standards. Some of these judgments are referenced directly to the syllabus and are usually detailed for the accreditation of school Work Programs and for certification of student performance at the end of a course of study.

However the vast majority of teacher judgments about assessment, take place in a classroom context. It is within this context that assessment serves its true educational purpose. The information about student performance which is profiled over time is informed by the intimate knowledge which teachers have of student achievement. This information is a summary of student achievement and reflects the interim standards attained. These interim standards then inform a teacher judgment about the attainment of global criteria at exit from a course of study. The Level of Achievement subsequently awarded represents the end of the pathway of teacher judgments — a pathway unique to each classroom.

## References


Sadler, D. R., *ROSBA's Family Connections*, Discussion Paper 1, Board of Secondary School Studies, January, 1986.

Sadler, D. R., *The Case for Explicitly Stated Standards*, Discussion Paper 2, Board of Secondary School Studies, January, 1986.

McMeniman, M., *A Standards Schema, Discussion Paper 3*, Board of Secondary School Studies, January, 1986.

Sadler, D. R., *Defining Achievement Levels, Discussion Paper 4*, Board of Secondary School Studies, January 1986.

Board of Secondary School Studies, *Handbook of Procedures for Accreditation and Certification*, March, 1985.

McMeniman, M., *Formative and Summative Assessment — A Complementary Approach*, Discussion Paper 6, Board of Secondary School Studies, January, 1986.

Sadler, D. R., *General Principles for Organising Criteria*, Discussion Paper 9, Board of Secondary School Studies, April, 1986.

Sadler, D. R., *Defining and Achieving Comparability of Assessment*, Discussion Paper 12, Board of Secondary School Studies, September, 1986.

# Assessment of Laboratory Performance in Science Classrooms

## Discussion Paper 17

**Abstract**: Laboratory activity in high school science classrooms serves a variety of purposes including psychomotor skill development and concept introduction and amplification. However it does not by itself provide sufficient experience that the learning outcomes are direct in the case of concept development, or of a high order in the case of psychomotor skill. This paper sets out a schema of global outcomes which might provide a more realistic framework for decisions about students laboratory performance at the end of a course of study. The use of the schema within the total course of study to award an exit Level of Achievement is also discussed.

**Authors**: Warren Beasley, Assessment Unit, and Peter Stannard, Burnside State High School, January 1987

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Board of Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. It should not be taken as representing Board policy, and is being circulated within the teaching profession to obtain reactions to its contents.



## Introduction

Laboratory work is generally regarded as a necessary and highly significant component of the science education experience provided for secondary school students. The list of desired outcomes produced by the Australian Science Teachers Association, ASTA, (1980) is representative of the international interest in this domain of science learning. The expectation is that science education should be concerned with developing skills in:

- making observations and measurements, both with unaided senses and instruments general manual dexterity
- the use of specialised equipment
- extracting, organising and analysing information from a variety of sources designing controlled experiments to test hypotheses
- thinking critically and creatively about ideas, information and problems
- interpreting and using mathematical data and methods in science communicating ideas related to science using a variety of media and methods.

The ASTA position is one which reflects not only the psychomotor domain of learning but incorporates the integrated process-skills located in the cognitive domain. However, the meaning of the world 'skill' in the above classification goes beyond that normally associated with 'skill objectives' in the ROSBA syllabus documents in science.

For example, the junior science Syllabus (1982) limits the use of skill to:

(a) Working with chemical and equipment: safe handling procedures, correct and economical usage and maintenance, and a safety conscious attitude;

(b) working with live specimens: correct handling of living and preserved materials;

(c) Fieldwork: developing field skills.

In the Chemistry Syllabus (1983), the following constitute skills. The student should be able to

(a) select, assemble, operate, dismantle, clean and return equipment;

(b) match equipment to the task;

(c) accurately read measuring instruments;

(d) handle, prepare and be aware of the hazards of chemicals;

(e) detect and eliminate errors in the setting up of equipment; and

(f) draw accurate diagrams of equipment.

The syllabuses in science have clearly identified the 'skills' domain as a legitimate area for student development and assessment. Although mentioning a small connection with the affective and cognitive domains, the syllabuses project a view of laboratory activity as being singularly important for the assessment of psychomotor domain. It does not take into account the wider purpose of laboratory work in science classrooms which has substantial cognitive and affective dimensions as well.

This paper attempts to place in perspective the nature of the laboratory experience in high school science classrooms (Years 8–12) and to derive the global outcomes of laboratory experience which could realistically be used when awarding an exit Level of Achievement in science subjects.

## Patterns of Laboratory Activity

Although there are variations of the pattern, Newman (1985) has described the structure of traditional laboratory works as follows.

> After a brief introduction by the teacher explaining what has to be done and how it has to be done, the student operates upon some materials and equipment to bring about some transformation of the same. The student may then have to process certain data available before, during and after the transformation process, ordering it, treating it mathematically, etc. During the hands-on activity and/or at its conclusion the student is prompted to communicate ideas related to the activity with other students and perhaps the teacher, and is often required to commit some material to writing. At the conclusion of the activity the teacher works with the class as a whole, attempting to have the students think about what has been done, what has been observed, calculated, etc. and its significance.

Such an activity sequence involves the student, or groups of students, in considerable manipulative activity and is accompanied by data processing and writing. Given the nature of traditional laboratory work, it is unlikely that many of the concepts, principles and theories of science will be acquired as a direct result of the laboratory experience. As well, the level of psychomotor skill development which is possible in the time available, is also constrained. However, considerable time, energy and expense are committed to laboratory activity in science programs. therefore it would be reasonable to expect this laboratory activity to contribute information towards a student's achievement in science. Taking these factors into account, what aspects of laboratory experience could be useful in determining the standards for assessing a student's laboratory performance? Stannard (1982) has outlined a range of teaching/learning settings which may be structured in a teaching laboratory. Table 1 provides four degrees of commitment to laboratory teaching which may suit a range of circumstances currently existing in science classrooms.

**Table 1: Four approaches to the structuring of laboratory activity.**

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| STUDENT ACTIONS TO BE ASSESSED | Experimental execution only | Experimental execution + Observations | Experimental execution + Observations + Interpretation and Explanation of data | Planning and design + Experimental execution + Observations + Interpretation and Explanation of data |
| TEACHER ACTION | Provide detailed stepwise procedure + Supply only necessary equipment to student | Provide detailed stepwise procedure + Supply only necessary equipment to student | Provide detailed stepwise procedure + Supply only necessary equipment to student | Give directions for investigation (cue notes) + Supply equipment at central point(s) |
| STUDENT WRITTEN RESPONSE | none | raw data recorded | raw data recorded + conclusions made and/or graphs drawn and/or mathematical relationships made | Procedures are explained + Raw data gathered + Conclusions made and/or graphs drawn and/or relationships made |

The circumstances surrounding each of the four approaches will vary, and are dependent upon:

(a) the time available in the laboratory

(b) the equipment and materials available

(c) the expertise of the teacher and laboratory assistant

(d) the number of students in the group

(e) the length of experience of students within the course.

This framework details a range of approaches to structuring the laboratory experiences of students. It has the distinct advantage of recognizing that both the psychomotor and cognitive domains of learning are necessary to complete a high school laboratory investigation in a meaningful way. It provides a continuum of expectations of the student, beginning with the execution of simple motor skills and progresses through various stages of increasingly sophisticated process skills which would be required to complete a scientific experiment.

Collectively these approaches exhibit what Kempa and Ward (1975) have described as the four major phases of student responsibility within the context of laboratory activity.

These are:

1. Planning and design of an investigation.
2. Experimental execution involving manipulative skills.
3. Observation of changes.
4. Interpretation and explanation of data.

The position which has been argued is one which accepts student laboratory activity as a purposeful and important component of science education with strong integrated contributions from the psychomotor and cognitive domains. To label the outcomes under the banners of ‘skill’, ‘process’ or ‘content’ would, therefore, seem to impose an artificial division of expected learning outcomes. Within the context of school laboratory activity, each does not exist in its own right but rather is an integral part of the overall process of scientific investigation of natural phenomena. At all year levels, it is this investigatory aspect incorporating cognitive and psychomotor skills which distinguishes the laboratory activity of students. The conditions under which such activity is structured for students does vary and this variance is recognised in Table 1.

## A Standard Schema

The analysis provided above, of the nature of laboratory activities in science classrooms defines the criteria and standards by which student laboratory performance could be assessed. The criteria are representative of the four degrees of student responsibility which constitute the possible dimensions of the laboratory investigations undertaken by students over a course of study. The schema (Appendix I) reflects a holistic view of the purposes and value of laboratory activities in science classrooms. As such it is not a mere listing of specific objectives for the course.

## Role of the Standards Schema

This schema (Appendix I) presents possible standards in laboratory performance to which science students could aspire. The cells of the schema are representative of a level of performance against each of the four criteria. These performance characteristics are intended to be general indicators of student performance at the end of a course of study in science subjects (i.e. Chemistry, Physics, etc.), and reflect a composite picture of performances on individual laboratory activities. However, one important qualification is needed. School-based assessment under ROSBA encourages consensus between schools and review panels about how a school decides to structure its course to achieve the aims of the syllabus. Circumstances within a school may be such that a criterion is not represented in a laboratory activity. For example, the criterion of Planning and Design may be negated if laboratory manuals present all the details for students and their learning experiences are limited to those described in the manual.

This paper does not attempt to provide the detailed mechanism by which laboratory performance over time could be monitored. It emphasises the nature of the laboratory experience in schools and establishes standards of global performance in this domain of the whole course of study. Information from such a laboratory performance schema and a ‘cognitive’ standards schema in each of the science subject areas together would provide the basis for awarding an Exit Level of Achievement. This latter schema would reflect the inherent characteristics of that subject and would be concerned more directly with the cognitive areas of the syllabus, presently defined as ‘content’ and ‘process’.

A method of combining information to derive exit Levels of Achievement is presented in Discussion Paper 7: Mathematics Criteria for Awarding Levels of Achievement (Findlay, 1986). A similar procedure is considered appropriate in the science subject areas.

## The Standard Schema: its relationship to the award of exit Levels of Achievement

It is essential that the central purpose of the standards schema be kept in mind. It represents a spectrum of global outcomes of student laboratory learning which would be possible in a course of study in a science subject. It does not represent a device for assessing single laboratory activities. Over time, teachers will observe students operating within the laboratory context. For criteria C1, C3, and C4, the information which is profiled will be gleaned from student practical note books, research reports, and field study reports.

Stannard (1982) has reviewed a number of techniques across science subjects which could be used to profile information on student performance relating to Criterion C2, Experimental Execution.

Levels of Achievement under ROSBA are intended to be representative of a student's ability at exit from a course of study. It is generally accepted that student outcomes in the area of laboratory performance are developmental in nature. The award of the exit Level of Achievement should reflect this growth of student learnings. The standards schema being proposed in this paper acts essentially as a mirror for the student achievement profile.

It is the teacher judgment of the congruence of the student profile and the specified standards from the schema which determines the exit Level of Achievement. This is illustrated further in Table 2.

**Table 2. Sample standards for awarding a Level of Achievement.**

| STANDARDS → | S4 | S3 | S2 | S1 |
|---|---|---|---|---|
| CRITERIA ↘ | | | | |
| C1 | | | X | |
| C2 | | | X | |
| C3 | | | X | |
| C4 | | | X | |

For example, a 'Sound Achievement' in science could be defined to include the set of standards in laboratory performance indicated by X above. Further discussion of the issues surrounding the definitions of Levels of Achievement and Standards Schema are provided respectively in Discussion Paper 4 (Sadler, 1986) and Discussion Paper 3 (McMeniman, 1986).

This paper has proposed a different focus on the assessment of laboratory skills in science. It has proposed that skills be taught and assessed in a similar context. This context is the laboratory investigation which students perform repeatedly throughout a course of study. It contains the dimensions of planning, execution, observation, interpretation and explanation.

A Standards Schema containing these broad dimensions has been proposed. This schema represents part of the mirror against which a student achievement profile is reflected in order that a teacher judgment can be made about the appropriate exit level of Achievement.

## References


Australian Science Teachers Association. 'Policy on Science Curriculum K–12', *Australian Science Teachers Journal*, 26, 3, 1980.

Findlay, J., *Mathematics Criteria for Awarding Exit Levels of Achievement*, Discussion Paper 7, Board of Secondary School Studies, January, 1986.

Kempa, R.F. and Ward, J.E., 'The Effect of Different Modes of Task Orientation on Observational Attainment in Practical Chemistry', *Journal of Research in Science Teaching*, 12, 1, 1975.

McMeniman, M., *A Standards Schema*, Discussion Paper 3, Board of Secondary School Studies, January, 1986.

Newman, B., 'Realistic Expectations for Traditional Laboratory Work', *Research in Science Education*, 15, 1985.

Sadler, D. R., *Defining Achievement levels*, Discussion Paper 4, Board of Secondary School Studies, January, 1986.

Stannard, P., 'Evaluating Laboratory Performanc'e, *The Queensland Science Teacher*, November, 1982.

# Appendix: A standards schema for the assessment of student laboratory performance

| STANDARDS → / CRITERIA ↓ | *S4* | *S3* | *S2* | *S1* |
|---|---|---|---|---|
| Planning and Design C1 | Presents a plan for a properly controlled experiment and discusses it critically. | Presents a plan which needs modification. Understands overall approach to the problem but some omissions in a critical discussion. | Presents a plan which is satisfactory but needs further detail. Shows little critical insight into the problem. | Presents a poor plan and shows little evidence of insight into how to translate the problem into action. |
| Experimental Execution C2 | Demonstrates a consistent ability to carry out the experiment work and takes account of the precision of the apparatus. | Demonstrates a consistent ability to carry out the experiment but is limited in a number of psychomotor skills. | Sets up the apparatus consistently but often needs advice to complete the investigation. | Careless in handling apparatus. Often fails to follow instructions. Consistently needs advice. |
| Observation of changes C3 | Correct observations specified, unexpected results recorded, errors are identified and explained. | Presentation of data consistent with experimental conditions, but little attention to errors. Observations lacking in fine detail. | Some consistency in presentation of data, but generally consistent with experimental data. Some measurements outside the range of the instrument accepted. | Poor presentation. Help needed in measuring. Poor discrimination ability in observing. No concept of error range. |
| Interpretation and Explanation C4 | Demonstrates a definite analytical approach, calculates results accurately, relates investigation to problem stated. | Presents an adequate analysis, calculates correctly, but liable to minor errors. Makes tentative judgments about problems stated and the investigation. | With assistance understands the data and proceeds with calculations. Selects appropriate steps in translating data into acceptable results. | Demonstrates little understanding of the relevant data and proceeds with difficulty in carrying out the calculations. |

# Profiling Student Achievement

## Discussion Paper 18

**Abstract**: The record of student performance over a course of study provides a summary of information from which ultimately a judgment is made by a teacher on an appropriate exit Level of Achievement. This summary, determined from and including qualitative and/or quantitative statements of performance, captures sufficient information to indicate standards of achievement. The judgements of standards attained are referenced initially to interim criteria at the end of semester, and finally to global criteria at the completion of a course of study. The design characteristics of a format to profile records of student performance are discussed.

**Author**: Warren Beasley, Assessment Unit, March 1987

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Board of Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. It should not be taken as representing Board policy, and is being circulated within the teaching profession to obtain reactions to its contents.



## Introduction

Tracking the performance characteristics of students is an essential task in any school-based assessment system, especially one which incorporates the principle of selective updating using the fullest and latest information about student performance. For a system which requires teacher judgements to be made with reference to criteria and standards, the method of recording assessment data poses significant problems of design and interpretation. This paper explores the nature of assessment data and its recording for use in the judgmental process of determining an exit Level of Achievement. It represents an extension of the ideas and issues raised in two previous Discussion Papers:

(1) *General Principles for Organising Criteria* (Sadler, 1986), and

(2) *A Pathway of Teacher Judgments: From Syllabus to Level of Achievement* (Beasley, 1986),

and creates a bridge between the principles of criteria-based assessment and its day-to-day practice..

## Objectives, Learning Experiences and Test Instruments

Within an accredited Work Program, objectives are separated in a useful but somewhat artificial way and labelled content, process, skill and affect. These objectives provide an emphasis or skeleton for teachers to plan and develop a series of lessons. The enactment of these lessons in the classroom is such that students see the content developing in an integrated and meaningful way.

A series of lessons provides a variety of learning settings to which students and teachers alike make a unique contribution. For the students, the interactional patterns provide an opportunity to learn about a topic in a particular subject. For example, a lesson sequence in geography may be perceived by students to focus on 'landscape patterns of the Darling Downs region'. For the teacher, the initial lessons in this topic could be focused more specifically on the affective domain, e.g. giving students a feeling of what it would be like to live in such an area. This might also include description of where the area is relative to other major geographical features. Overlapping of content, process, skill and affective objectives is a natural consequence of the teaching and learning in any topic. The integration of the objectives within a topic gives meaning to the individual objectives. The presentation of the topic requires the simultaneous development of a variety of objectives.

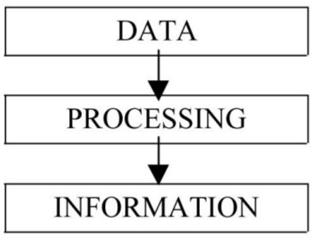
Formal assessment is based on an appropriate assessment instrument. The design specifications for the instrument are intended to provide a context which is consistent with the way the subject matter was taught and learned. As emphasized in the previous paragraph, objectives are not learned in isolation from the topic. For example, to address the proposition ‘riverflat soils provide a suitable medium for the production of staple crops’, it would be more meaningful for students to develop an understanding of this proposition within a broad context. The topic area ‘landscape patterns of the Darling Downs area’ provides this broader context and this proposition is taught and learned along with the simultaneous development of the related concepts.

Individual objectives should not be interpreted in isolation, but should be seen to be derived from the more salient dimensions of the course of study. These dimensions provide a focus towards which student performance should be shaped. A particular selection of content, process and skill objectives from a Work Program may appear to be balanced but in reality the selection may be quite arbitrary. These classification labels do not represent the focus by which subjects are taught and learned in the classroom or represent a true reflection of the subject. Criteria which are derived from the salient dimensions of the subject can be interpreted by students and thereby help in identifying present strengths and weaknesses in the subject. Students will interpret performances in the same context in which the topic was taught in the classroom, and hopefully move toward self-directed improvement.

Subjects exist in the school curriculum because of the contribution each can make to the balanced education of students. The dimensions for each subject are described in syllabus documents and therefore any assessment program must have these dimensions as its focus.

## Steps in Profiling Performance

The general pattern of the events which characterises the assessment of student performance can be represented thus:

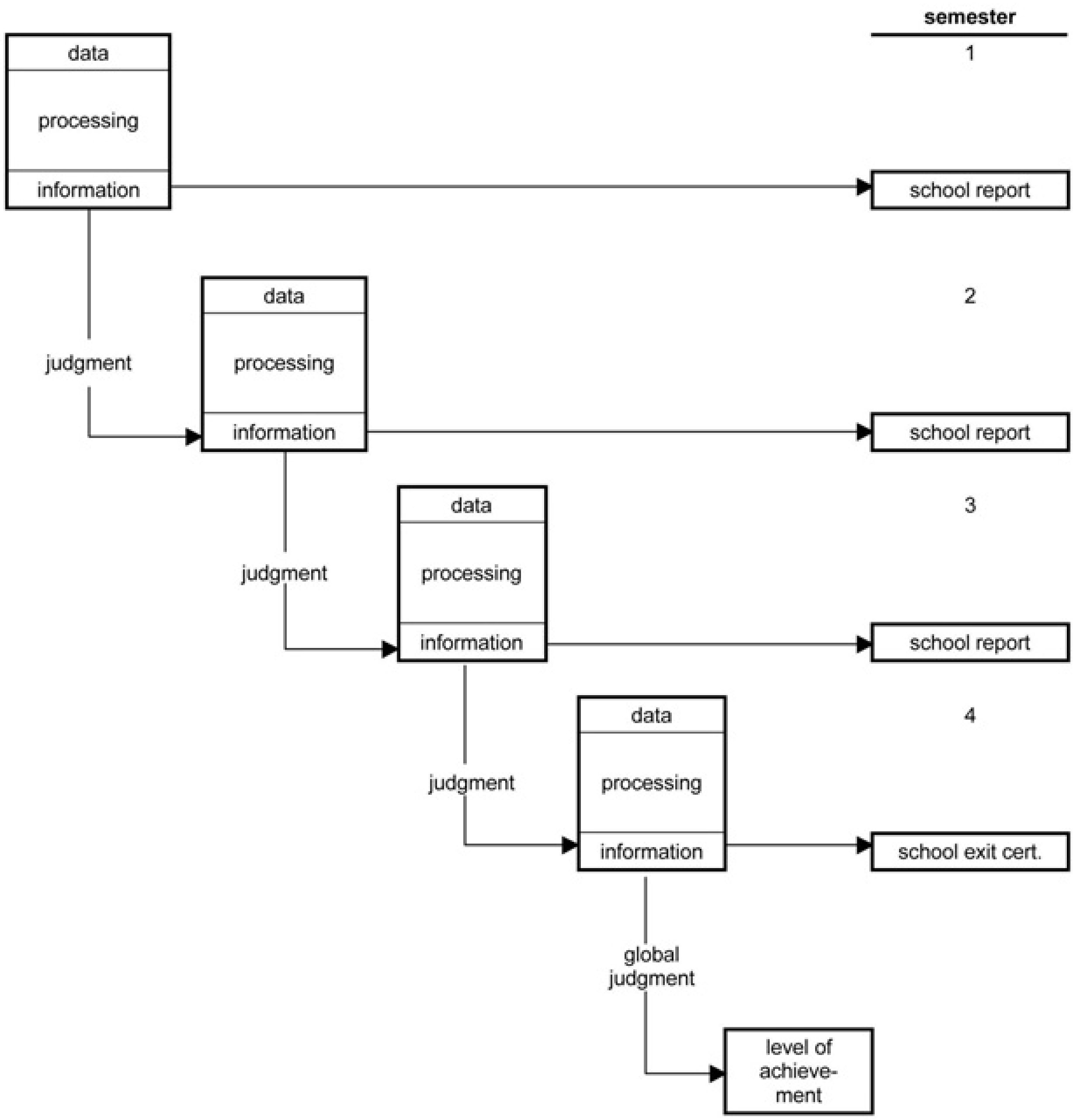



Data are the facts about student performance as derived from the assessment instrument. These data are usually referenced to performance criteria which have been promulgated in advance to students before the instrument is administered.

Information could be defined as the derived meaning of the consolidated data and informs the student of current performance characteristics. It should be a qualitative description of the levels of student performance at that time. For example, the properties of a student script (the data) contributing to the Written English Assessment program could be referenced to Level 2 criteria such as Content, Organisation, Language and Presentation. A qualitative statement about student performance in Written English which provides information of the current achievement status of the student is recorded (The Information). (See Discussion Paper 9, General Principles for Organising Criteria, Sadler, 1986). The sequence repeats itself throughout the course of study.

The realities of school practice often dictate that interim information about students be reported to parents and students at least once per semester. Where a student is not completing a course of study during that reporting phase, the performance characteristics are referenced to interim criteria designed by the school which are derived from and fully consistent with the global criteria expressed in a syllabus. For the majority of students in the senior school, the normal expectation would be to complete a course of study over four semesters. Therefore, for the teacher, the fabric of decision-making would be sequenced in the following way.

**Figure 1. The sequence of teacher decisions in reaching an exit Level of Achievement.**




The pattern of teacher judgments, in which information about student performance from one semester is carried forward and reinterpreted in light of new data generated during the subsequent semester, is cumulative and updated. The format in which data are recorded needs careful consideration. The actual ‘symbol’ which is recorded must have meaning in its own right, that is, the ‘symbol’ is referenced to criteria at Level 2 or Level 1 (Sadler, 1986).

It has been common practice to code achievement by means of a number or as a group of numbers. For example, content, lower order process, higher order process and skill categories find favour in science subjects which have a significant quantitative dimension, and 'marks' in each category are recorded directly from the completed test instrument. To treat these 'marks' as some form of currency, and to add weighted totals over four semesters to produce a consolidated group of numbers is foreign to the notion of using criteria and standards as the basis for awarding achievement levels. Numbers need to have meaning and should be little more than a code referenced to qualitative statements of interim criteria which reflect the intended outcomes of a course of study.

It is the encoding of the 'symbolic data' in any semester which is carried forward as information to the following semester. This information should:

(i) differentiate between development and unitized aspects of the course;

(ii) indicate a current standard of performance using interim criteria: and

(iii) reference the global standards from which these interim standards have been derived. These should be accepted state-wide standards promulgated through syllabus statements.

This information base becomes a point of reference for new assessment data generated over the next semester. The new data will allow selective updating of the information on student performance to be carried forward. Data which are representative of developmental aspects of a course are selectively updated in light of emerging characteristics. Data which are unitized (i.e. referenced to criteria which apply only to that section of the course) are carried forward to the end of the course and then interpreted in terms of global criteria.

Finally, by the end of a course of study a profile of information about student performance is generated. This profile contains an interpretation of the standards reached by students on the nominated criteria. It is this information which is compared with the standards schema promulgated through the syllabus, and a judgment made as to the appropriate exit Level of Achievement.

## Recording the Profile

The conceptualisation of a scheme of continuous but not cumulative assessment culminating in an exit Level of Achievement using a notion of criteria and standards has been summarised in previous sections. However, the logistics of implementing such a scheme at the class room level are not so apparent.

Any system for recording assessment data is by necessity a summary of teacher judgments about student performance. Within the ROSBA context, the recording system needs to satisfy a number of requirements. The profile should allow for:

(i) a description of the assessment instrument;

(ii) a coding of assessment data;

(iii) an interpretation of data in terms of criteria met;

(iv) selective updating of student performance characteristics across semesters; and

(v) descriptive statements of student achievements to be produced (for reporting to parents).

One method of structuring a student profile is presented in Figure 2. This system emphasises the importance of teacher judgment in the assessment process. The information cells represent the ongoing and updated judgments of student performance which are taken on board during the following semester. Preferably, it should be a qualitative description of standards reached and be updated at the end of each semester in the light of the new assessment data generated.

It is the information at the end of each semester which provides the basis for periodic reporting to parents. In semester 4, an exit Level of Achievement, referenced to the global criteria and standards in the syllabus document, is generated for the purpose of Board certification.

The particular example provided in Figure 2 comes from mathematics. It takes the standards schema provided in Discussion Paper 7: Mathematics Criteria for Awarding Exit Levels of Achievement (Findlay, 1986) for descriptions and coding of appropriate standards (S1 to S6)

The criteria for assessing student performances are:

| | |
|---|---|
| Knowledge | (K) |
| Representation | (R) |
| Manipulation | (M) |
| Procedure | (P) |
| Attack Mode | (AM) |

In the interpretation of the standards recorded for each semester, it should be noted that these interim standards are referenced to exit levels in the schema. The information which is profiled in each semester indicates whether the standards reached are reasonable for this stage of the course for that student.

In the final semester, teacher judgment of the exit standards on each criteria are noted, and takes account of Board policy on exit assessment (fullest and latest information, etc). The combination of standards attained in this example are representative of an award of High Achievement.

## Figure 2. A representation of the elements required in recording student performance over time

*SUBJECT: MATHEMATICS* *NAME: STUART ANDREWS* *STUDENT PROFILE*

**Semester 1: Prep Maths** — Data sem 1

| *Assessment instruments* | *K* | *R* | *P* | *M* | *AM* |
|---|---|---|---|---|---|
| *A*. calculator proficiency | S4 | – | P | – | – |
| *B*. Graphing project | – | R | P | M | – |
| *C*. Written test: basic skills | K | – | P | M | – |
| *D*. problem-solving assignment | – | – | – | – | S4 |

Semester 2: Algebra and Calculus 1 — **Data sem 2**

| Criteria | *K* | *R* | *P* | *M* | *AM* |
|---|---|---|---|---|---|
| E. Written test 1 | X | S2 | S2 | S2 | – |
| F. Written test 2 | S1 | S2 | S2 | S2 | S3 |

Semester 3: Geometry and Calculus 2 — **Data sem 3**

| Criteria | *K* | *R* | *P* | *M* | *AM* |
|---|---|---|---|---|---|
| G. Graphing project | S4 | S4 | S4 | S3 | S4 |
| H. Written test | S4 | S4 | S5 | S4 | S4 |

Semester 4: Statistics and Probability — **Data sem 4**

| Criteria | *K* | *R* | *P* | *M* | *AM* |
|---|---|---|---|---|---|
| I. Calculator proficiency | S5 | – | S5 | – | – |
| J. Assignment (o, r, etc) | S4 | – | S4 | S4 | S5 |
| K. Written skills test | S4 | S4 | S4 | S4 | S4 |

*Information sem. 1**

Stuart: you are achieving the highest standard as presented in the assessment. Results are given only for the criteria which are the main focus of assessment.

*Information sem. 2**

Stuart: you have not learnt some of your basic procedures associated with max/min. and this affected your performance in all of the other criteria during the semester.

*Information sem. 3**

Your basic facts and skills have developed very well. The effect on your performanceon other criteria has been substantial. You are performing at the highest standards provided in the assesment instruments during the semester, and this semester's result will affect your exit level substantially.

*Info. sem. 4**

Interpretation in assignment and written test was well handled. Consistent evidence of good problem-solving ability. Knowledge of basic skills over all the course has developed substantially during this year of the course.

**Fullest & latest information**

K: S4
R: S4
P: S4
M: S4
AM: S4

INFORMATION STATEMENTS TAKE ACCOUNT OF PREVIOUS SEMESTER PERFORMANCE

**Key**

– not major focus
X has not achieved lowest standard tested

**Level of Achievement: HA**

## References


Beasley, W.F.: 1986. *A Pathway of Teacher Judgments: From Syllabus to Level of Achievement*, Discussion Paper 16, Board of Secondary School Studies.

Findlay, J.G: 1986. *Mathematics Criteria for Awarding Exit Levels of Achievement*, Discussion Paper 7, Board of Secondary School Studies.

Sadler, D.R; 1986. *General Principles for Organising Criteria*, Discussion Paper 9, Board of Secondary School Studies.

# Principles for Determining Exit Assessment

## Discussion Paper 19

**Abstract:** Six key principles underpin exit assessment. They are continuous assessment, balance, mandatory aspects of the syllabus and significant aspects of the course of study, selective updating, and fullest and latest information. These principles are explained in some detail and related to courses of study.

**Author:** Jan Findlay, Assessment Unit, March 1987

***Note:*** This Discussion Paper has been produced by a member of the Assessment Unit of the Board of Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. It should not be taken as representing Board policy, and is being circulated within the teaching profession to obtain reactions to its contents.



### Introduction

School-based curriculum development and assessment are integral to both assessment systems which have operated in Queensland over approximately the past 15 years. While school-based systems in Queensland operate with substantial school autonomy, the co-existence of a central accreditation and certification authority (Board of Secondary School Studies) with school-based assessment implies that the autonomy is not an unlimited quantity (Sadler, 1986). The requirement to satisfy the six principles of exit assessment defines some of the bounds of school autonomy, albeit indirectly. This paper considers these principles in some detail as a general introduction to, and background information for, the particular syllabus interpretations.

### Certification

Board certification is possible for those subjects within a school for which work programs have been accredited. Accreditation results when work programs satisfy the Form R2 checklist against which they are compared. The checklist concerns such elements as specific objectives to be pursued, examples of learning experiences and the assessment program (Items 8 to 12). The Form R2 for Board-Registered School Subjects varies slightly in items 1–7 from the corresponding Form R2 for Board Subjects. However, the items referring to the assessment program are the same for each.

Within the assessment program, schools are required to indicate the summative assessment plan, the variety of techniques to be used, the criteria used and the rules for combining performances on these criteria to obtain the exit achievement levels. The descriptions of such elements of the assessment program provide information about implementation of the principles of continuous assessment, balance, inclusion of mandatory aspects of the syllabus, significant aspects of the course of study, selective updating and fullest and latest information, in determining the exit assessment. The Form R2 as a tool for accreditation therefore influences the designed curriculum. Since accreditation is a prerequisite for certification and certification uses the work program as a reference, the Form R2 may also be perceived to be indirectly connected with certification. However, the main conditions to be fulfilled for certification occur at later stages. Supply to the Board by schools of proposed numbers of students in each achievement level, along with samples of student work representing achievement at each level are two of these later conditions to be satisfied in certification procedures.

While highly supportive of curriculum development, such as schools writing work programs for Board Subjects, and developing Board-Registered School Subjects and School Subjects, the Board stated its major concern of *exit assessment* and certification in a letter from the Executive Officer to all teachers in October, 1984. In clarifying this focus, the letter delineated the principles that must underpin exit assessment if it is to *reflect achievement in the course of study up to the date of exit*. The principles state that:

1. Information is to be gathered through a process of continuous assessment.
2. Balance of assessments is a balance over the course of study and not necessarily a balance within a semester or between semesters.
3. Exit achievement levels are to be devised from student achievement in all areas identified in the syllabus as being mandatory.
4. Assessment of a student's achievement is to be in the significant aspects of the course of study identified in the syllabus and the school's work program.
5. A student's profile of achievement is to be selectively updated.
6. Exit assessment is to be devised to provide the fullest and latest information on a student's achievement in the course of study.

One of the five labels of Very High Achievement, High Achievement, Sound Achievement, Limited Achievement, and Very Limited Achievement is used to express a student's exit assessment.

Interpreting the six principles of exit assessment is essential to the implementation of them within the assessment program, and, indirectly, to the entire teaching program of a course of study. The interdependence of the six principles become evident as each is interpreted. Before considering these principles however, a discussion of courses of study will be presented. The structure of the course determines how the principles, especially selective updating and fullest and latest information, are implemented to obtain information from which exit Levels of Achievement are determined.

## Courses of Study

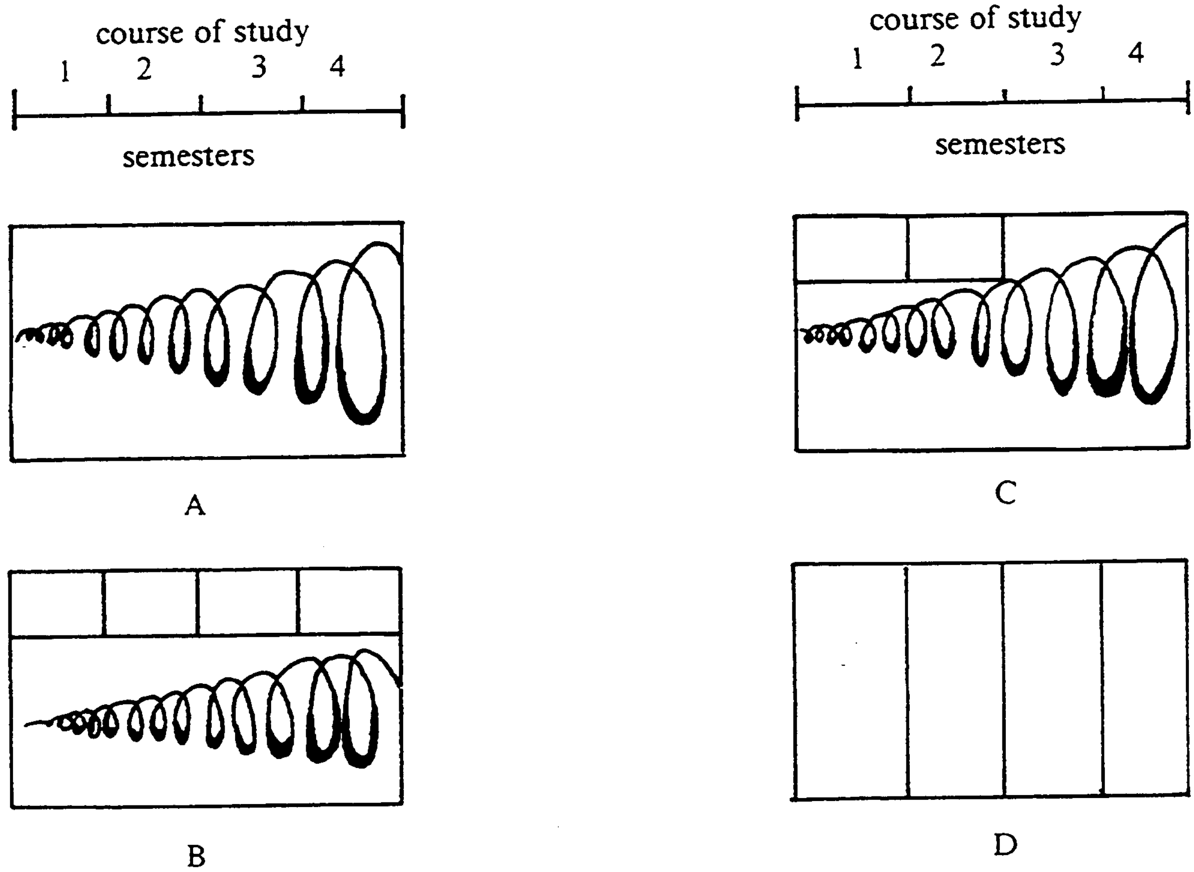
A number of factors must be considered when designing a course of study and writing the associated work program. A subject consists of various components. Mathematics, for example, is perceived to be both a process and a body of knowledge about concepts and procedures from a variety of fields within it. English is considered by some to be purely a 'process-oriented' subject, where content essentially is immaterial. A course of study reflects the important components of the subjects, their relative emphases, the degree of integration of the components, as well as their developmental or non-developmental character, and the effect this developmental feature has on the sequencing of the learning experience. Knowledge of the ways children learn influences the structure of a course of study as well.

It is customary to use a spiralling curriculum approach on developmental components of a subject. These components are revisited constantly throughout the course, and at each level, treated in a more sophisticated manner. In this approach, new aspects introduced into the components 'add-on' and allow the ever increasing spiral to generate. Where components are not considered developmental, the level of sophistication of treatment remains unchanged even if revisiting occurs. Generally, these types of components are taught and learned within a modular or compartmentalised curriculum structure. Within many of the existing Syllabuses, 'skill' and most if not all 'process' objectives are perceived to be developmental in nature, with 'content' objectives non-developmental in character.

Many syllabuses include a selection of semester units of work. In some cases, these semester units do represent distinct, non-developmental modules of work. In other cases however, the semester unit of work represents a conveniently sized section of the developmental course labelled as a particular unit. Some syllabuses do not divide work into semester units. For these syllabuses, the semester unit in the work program merely reflects the structuring of the school calendar year and its effect on courses of study.

The combination of all these factors results in a diversity of models of courses of study, and, ultimately, ways in which fullest and latest information is obtained through selective updating. The selection of models in Figure 1 is not meant to be exhaustive, but merely to exemplify some of the present practices and structures within the system.

### Figure 1. Various models of course of study over 4 semesters.




 represents spiral curriculum approach and implies developmental nature.

### Model A

In this model, all components of the course of study are organised using a spiral curriculum approach, as they (and the subject itself) are perceived to be totally developmental in nature.

### Model B

This model indicates that some components, or parts of components are developmental in nature, whereas other components or parts thereof are non-developmental. Consider, for example, a simple model in which there exist two components, A and B. The following situations are possible.

- One of the components is entirely developmental while the other component is not.
- One component is entirely developmental, while the second component has both developmental and non-developmental elements.
- Both components have elements which are developmental as well as elements which are non-developmental.

The course of study treats developmental parts using a spiral curriculum approach, and the non-developmental aspects in a modular fashion. The developmental parts allow the various semester units of work to be linked, and in this way the effects of modularising the work are diminished.

### Model C

This model is a composite of models A and B. It implies that the initial two semester units of work follow a pattern of model B, but then during the later semesters the work presented is completely developmental and builds on the developmental components of the first two semester units. A variation of model C might be one in which the modular aspects occur during the last two semesters.

### Model D

In this model, the components *over a course of study* are not developmental and are presented as distinct modules (in this case, semester units). The developmental character of components within a module may be a possibility, but this aspect is not considered here.

## Interpreting the Six Principles

### Continuous Assessment

The major operating principle is ‘continuous assessment’. The process of continuous assessment provides the framework in which all the other five principles of balance, mandatory and significant aspects of the syllabus and of the course respectively, selective updating, and fullest and latest information exist and operate.

In discussing continuous assessment, the two terms ‘continuous’ and ‘assessment’ will be considered, one at a time. A common interpretation of the term ‘assessment’ is that it is the result which a student obtains on assessment instruments. Assessment is also the judging process performed by a teacher while reviewing students’ performances. Assessment therefore may be considered to have both process and product dimensions. By definition ‘continuous’ is synonymous with ‘no breaks’. Continuous assessment as a process is thus a constant judging of student achievement with an implication that the result or product is not a static quantity, that is, selective updating occurs. The results of the judging process in relation to stated criteria provide information which is used to determine the exit assessment.

Continuous assessment as a process is an inherent dimension of the daily teacher-learner interaction. It is the mechanism through which teachers gauge progress, understanding, timing of introduction of new concepts, and the many other factors which are all parts of the teaching-learning environment. It draws on the many processes that characterise the interaction. In the day-to-day teaching-learning context, teachers consciously and often unconsciously update their perceptions of students informally through observations, individual and small group discussions, quizzing, class questioning and other interactions even outside the classroom environment. The basic function of these unrecorded perceptions is for formative purposes. The formal dimension of the continuous assessment process is the source of information used for summative purposes. It occurs through planned formal assessment settings such as written tests, practical tests, assignments and orals, and complements teacher knowledge of students gained through the informal modes of assessment. The product of these two modes of obtaining information constitutes continuous assessment as a process which, by definition, must occur over a period of time.

Continuous assessment, as envisaged in this system, does not necessarily imply very frequent or continual formal assessment; nor is it perceived to be equivalent to cumulative or terminal assessment. In cumulative assessment, scores are merely aggregated throughout the course and the developmental nature of student learning is not accounted for. Exit assessment must satisfy concurrently the six principles associated with it. The principles of selective updating and fullest and latest information could not function effectively in cumulative assessment. Similarly, in terminal assessment (one final assessment), the principles of selective updating and fullest and latest information would not exist.

## Balance, and Mandatory and Significant Aspects

These three aspects, when referred to exit assessment, assume specific meanings in syllabuses. Common interpretations of mandatory in this context include compulsory minimum subject matter (core) or general objectives and/or components such as practical and theory areas to be included in exit assessment judgements. Significant aspects in this context generally refer to that extra subject matter or those extra objectives over and above the compulsory minimum ones, that are considered highly important in the assessment of students' learning. Mandatory aspects are the compulsory subset of significant aspects. Balance requires appropriate inclusion of the significant (and thus mandatory) aspects in the exit assessment, and relates to the course as a whole. Balance in the summative (exit) assessment generally reflects the amount of time spent on particular significant subject matter and objectives. It is, in effect, the matching of exit assessment with the course interpretation of significant aspects to be included. Balance has implications for the process of selective updating and thus fullest and latest information. In a developmental course of study, for example, the significant aspects of the course are constantly revisited, and developed. Balance in exit assessment here implies that information from the latter part of the course constitutes the fullest and latest information. This principle along with selective updating is explained more fully in the next section.

## Selective Updating and Fullest and Latest Information

These two principles operate within the context of continuous assessment. Selective updating is the process through which fullest and latest information about a student's achievement is obtained. Selective updating operates on data which are made redundant because more recent data absorb or supersede the previous data. The fact that 'fullest' and 'latest' may not be synonymous requires both aspects to be considered. Fullest and latest information may be obtained in two basic ways. Either the latest information supersedes all previous information and it is both the fullest and latest information, or alternatively, the fullest and latest information consists of both the most recent data on developmental aspects together with any previous but unsuperseded data (non-developmental course aspects). The inclusion of unsuperseded data from earlier in the course of study implies that it contains at least mandatory aspects of the course which must contribute to exit assessment.

The two principles when applied to each of the four models of courses of study presented earlier in this paper will serve to clarify their application. For model A, the exit assessment derives from information gathered during the latter part of the course. In this developmental subject, significant (and mandatory) aspects of the course are constantly revisited, and later data supersede earlier data. In the developmental-compartmentalised model B, the exit assessment for the developmental aspects of the course derive from the latter part of the course. Information for those non-developmental aspects containing mandatory and significant aspects must also contribute to exit assessment. In this model, fullest and latest information consists of the latest information from each module on non-developmental aspects, together with the latest information on developmental aspects obtained in the latter part of the course. For model C, the latest information for the developmental aspects (latter part of course) together with the latest information on non-developmental but significant aspects (obtained from the two modules) constitute fullest and latest information. In the last example, model D, fullest and latest information for exit assessment is obtained by considering the fullest and latest information from each distinct module. To obtain the fullest and latest information within a module, the structure of the module would need to be identified. It is possible that Model A, B or C could represent that internal structure, and once again this would influence how the fullest and latest information within the module is obtained.

This nexus, between the course structure and the method for deriving exit assessment using the principles of selective updating and fullest and latest information, deserves attention. For example, some work programs have been written to reflect the structure and properties as shown in one model but the assessment program is compatible with an alternative model. Logically, when this conflict arises, the assessment program should be reconstructed to reflect the course structure.

## Conclusion

Exit assessment incorporates six interrelated principles. Continuous assessment is the context in which the other five principles operate. Implementing these principles requires an understanding of course structures and properties, the salient feature of each exit assessment principle, their interdependence, and their relation to various courses of study.

## References

Letter to all teachers, (Board of Secondary School Studies, 1st October 1984) Exit Assessment, (Board of Secondary School Studies, Inservice summary document, 1986)

Sadler, D. R; 1986. *School-Based Assessment and School Autonomy*, Discussion Paper 11, Board of Secondary School Studies.

# Issues in Reporting Assessment

## Discussion paper 20

**Abstract**: This paper is about some of the issues associated with reporting in schools. Provided are a number of suggestions which may form partial solutions to problems which arise.

**Author**: Jan Findlay, Assessment Unit, March 1987

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Board of Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. It should not be taken as representing Board policy, and is being circulated within the teaching profession to obtain reactions to its contents.



### Introduction

A number of significant issues are associated with reporting by schools. However, the three issues considered in this paper are:

- achievement labels and their appropriate use;
- reporting times chosen by schools; and
- assessment for dual purposes.

The paper includes for each issue a number of suggestions for changes to practices now existent in some schools.

### Achievement labels and their appropriate use

Board policy stipulates that the official nomenclature of 'Very High Achievement' to 'Very Limited Achievement' be used in reporting exit student achievement on Board Certificates, as well as during the pre-exit monitoring and reviewing procedures for certification. The ROSBA recommendations suggest using these labels to report both overall student achievement and achievement on each of the broad criteria of the subject. The requirements of the Board coupled with the ROSBA recommendations have influenced some schools to perceive these labels as appropriate for all

reporting and recording processes in a school. Consequently, School Progress Reports and School Exit Certificates often express achievement for both the broad criteria of the subject and the overall (global) achievement, in terms of the official exit labels. In some schools the official labels are used also to report:

- global achievement for a particular module of work after it has been completed;
- global achievement on individual instruments; and
- achievement on individual assessment items.

For example, a student may have an assessment reported using one of the official labels for a mid-semester test, for an assignment, or for a unit test. In some cases, assessment instruments have been structured so that 'VLA' tasks are placed at the beginning of an assessment instrument, then 'LA' tasks, proceeding through to 'VHA' tasks at the end. Students are then classified according to their performance on various 'groups' of tasks on the instrument.

The use of the official labels for reporting student achievement for modules of work, assessment instruments and assessment items arises for a number of reasons. One reason relates to a lack of understanding of the underlying conditions that should be satisfied before the labels are used. These conditions are associated with summative assessment and it is implicit when a student is awarded a particular Level of Achievement that:

i. information has been collected over an extended time frame;

ii. the information has been collected from a number of instruments; and

iii. the resultant global Level of Achievement is a consequence of reviewing and judging these assessments *collectively*, against exit criteria and standards.

This relationship between the labels and the conditions for summative assessment has never been defined explicitly. Further, Board policy states only when use of the official labels is required, not the converse. In addition, there has been a tradition and expectation of collapsing information into concise forms, and schools naturally see the official labels as a ready-made form to use.

Some adverse consequences have arisen from the use of the official labels to report student achievement at exit and other times, achievement on individual criteria and global achievement, and achievement for modules of work, individual instruments or items.

Board syllabuses and work programs define the conditions for awarding the various exit Levels of Achievement. The meanings of the official categories from 'Very High Achievement' to 'Very Limited Achievement' when used in Progress Reports are not generally provided by the school as part of the Report. Students are left questioning whether the Levels of Achievement awarded in Progress Reports relate to the exit standards or to interim standards, and if the latter, what these interim standards are.

Another problem concerns the number of distinguishable standards for each criterion. There has been a tendency in some cases to create five standards of performance per criterion (whether or not five standards exist) and, further, label those standards VLA to VHA. The existence of five categories relating to global achievement does not imply that for <u>all</u> criteria of performance, there should normally be exactly five standards. Often where five standards (labeled VLA to VHA) on criteria are used, the one descriptor represents, for example, both standards of LA and SA.

The practice of using these official labels at interim times has fostered, for some students, unrealistic expectations in terms of the prediction of their exit Levels of Achievement, and often a belief that they have been stereotyped by the interim Levels of Achievement awarded. Use of the labels for individual instruments has also encouraged some people to believe that global achievement is obtained by some averaging process on the individual achievements, in a manner comparable to numerical averaging.

The common practice of using the official labels to report achievement on short modules of work, on individual tests, or on items, also dilutes the concept of summative assessment. This dilution of the concept occurs in these situations because a summative assessment judgment is no longer the result of comparing a number of assessments of a student's achievement collected over an extended time period, with set criteria and standards. That is, the time and/or quantity conditions for summative assessment may not be satisfied. Collapsing information about standards on a number of criteria into a single result for a particular test or module of work, has repercussions later in the derivation of a global Level of Achievement. This exit assessment is to be derived from a consideration of achievement on the broad criteria of the subject. The collapsing process runs the risk of submerging completely any visible connections between criteria and the global Level of Achievement, and pays only lip-service to criteria-based assessment.

Presented below are suggestions about practices which may help to avoid some of the above consequences.

**I**. Use the official labels to report global achievement only. Do not use the labels to report standards of achievement on the component criteria of global achievement or for modules of work, individual instruments and items. Use the labels to report exit global achievement only. However, this solution does not appear to take account of Board requirements associated with monitoring and reviewing practices, which require interim Levels of Achievement for students. It probably would not find great support among students and parents who expect and desire some indication of prospective exit performances, even when it is known that these prospective levels may not reflect the final performance.

**II**. An alternative is to use the labels to report global achievement only, but at both exit and at interim reporting times. In this case however, schools should include a rider in Progress Reports, which states that award of a certain interim Level of Achievement does not automatically guarantee an equivalent achievement level at exit. The rider might read as follows: 'This Level of Achievement is the best approximation that can be provided at this time based on the work covered and the present quality of performance by the student. However, award of a certain interim Level of Achievement does not guarantee the same Level of Achievement automatically at exit. This exit and interim Levels of Achievement may not always correspond because: (a) where subjects, or parts thereof, are developmental in nature, it is to be expected that exit standards are ordinarily more sophisticated (or at a higher level) than interim standards, and (b) an exit judgment takes into account all mandatory and other significant aspects of a course. At earlier times in the course when interim Levels of Achievement are awarded, not all mandatory and significant aspects of the course will have been studied.'

**III**. Record information on achievement in a subject against criteria, not against test instruments.

**IV**. When labelling achievement on each broad criterion of the subject, do not use VHA, HA,…Use an alternative coding such as $S_1$ to $S_n$ or A,B,C,…, (preferably not numbers), to describe the standards on the individual criteria. Ensure the codes have meaning. Such practices would release criteria from automatically having 'five standards'.

**V**. Publicise the rules for combining performances on individual criteria to obtain each global Level of Achievement. Such changes would help to clarify how the global achievement is obtained.

**VI**. Do not label individual items in test instruments as VLA, LA, SA,…as this is inconsistent with conditions i and ii associated with the labels; that is, the information is collected over time, from a number of assessments. Analyse and code items or groups of items in terms of standards on criteria.

## Reporting Times Chosen by Schools

For many schools, a general pattern for official progressive reporting in each year is as follows:

- near or at end of first term;
- end of first semester;
- end of second semester.

Of course, many departments within schools report unofficially on and to students at more frequent intervals. This unofficial reporting takes the form of formal feedback provided with individual assessment instruments and tasks, and the day-to-day feedback which is integral to the teaching–learning situation. Some school departments have introduced a 'semi-official reporting' scheme which requires students to show their parents results of assessments for a module of work and so on, and to obtain a parental signature and often comments on the work.

The ends of term or semester have been chosen by most schools as official reporting times for a number of reasons. The ROSBA recommendations concerned with reporting cite ends-of-semester as possible reporting times. Another factor influencing practice is the flexibility principle operating within the Senior school. This principle operates with the semester as the basic time unit and allows students to choose among Board Subjects, Board-Registered School Subjects, School Subjects and subjects studied at other institutions e.g. TAFE. Students are able to study the first three types of subjects for one to four semesters. Reporting times have traditionally been strongly associated with the timing of school holidays. The previous Radford system of assessment is another residual influence. It required official summative assessments at the ends of semesters (for the Senior school). This consolidated the position of the end of semester as a convenient reporting time for all sections of the school. The re-positioning of school holidays to coincide with end of semesters further entrenched these times as appropriate for reporting.

Discussion Paper 8, *Developing an Assessment Policy within a School* (Sadler, 1986), suggests that long-term benefits might accrue to both students and teachers if school reports were not so tightly coupled with the timing of tests. Staggering of official progressive reporting for different grades within a school, throughout the year and/or semester could provide one solution to the time problem which confronts teachers testing and reporting for all classes at the one time. Another alternative is staggering of official progressive reporting for different departments within a school, throughout each year and/or semester. The times chosen would reflect subject structure, (and the sequence of learning experiences over the year), needs of students and parents, and consideration of the other departments of the school. Severing the nexus between assessment and reporting times could help to equalise teachers' work loads and allow assessment to become a more natural element of teaching and learning.

## Assessment for Dual Purposes

The two purposes of assessment, formative and summative, are discussed in Discussion Paper 6, *Formative and Summative Assessment — A Complementary Approach* (McMeniman, 1986). The intended outcomes may arise through the use of separate instruments (often called 'formative' or 'summative' instruments), or they may evolve from the one instrument or task being used for the dual purposes. A common practice in some subjects presently is to provide little or no formative assessment to accompany 'summative instruments'. Educationally, such information is still valuable. Further, combining both purposes when assessing might lead to decreased formal testing for formative purposes, and more time for teaching and learning.

## Conclusion

Many issues are associated with reporting in schools. Some current practices connected with the three issues considered in this paper have potentially adverse consequences. These consequences are described, and alternate practices are suggested as a means of avoiding or mitigating the effect of these consequences.

## References

McMeniman, M; 1986. *Formative and Summative Assessment — A Complementary Approach*, (Discussion Paper 6), Board of Secondary School Studies.

Sadler, D. R.; 1986. *Developing and Assessment Policy within a School* (Discussion Paper 8), Board of Secondary School Studies.

Scott, E.; et al; 1978. *A Review of School Based Assessment in Queensland Secondary Schools*, Board of Secondary School Studies.

# The Place of Numerical Marks in Criteria-Based Assessment

## Discussion Paper 21

**Abstract**: Numerical marks form the currency for almost all assessments of student achievement in schools, and the use of them is rarely if ever challenged. In this paper, a number of assumptions underlying the use of marks are identified, and the appropriateness of marks in criteria-based assessment is examined. The conclusion is drawn that continued use of marks is more likely to hinder than to facilitate the practice of judging students achievements against fixed criteria and standards.

**Author**: Royce Sadler, Assessment Unit, March 1988

***Note***: This Discussion Paper has been produced by a member of the Assessment Unit of the Queensland Board of Senior Secondary School Studies. It is recommended to the teaching profession for consideration in the formulation of curriculum and assessment policy within secondary schools. Reactions to and comments on its contents would be helpful.



Although some authors use the term **mark** to refer to a teacher's comment of annotation written on a piece of work, or to tick (check) or cross, marks are taken here to mean the numbers or scores used to report or record the achievements of students. The term **grade** refers to a global summary award for a whole course or semester. Marks are in almost universal use in schools, and give rise to such phrases as test **scores**, 17 out of 20, an average of 53%, and "part marks" for problems not completely solved.

Marks are arrived at in a variety of ways, the principal three being forward counting, backward counting, and direct assignment. The first consists simply of counting the number of items correct. The items may be of different types, including sentences to be completed, equations to be balanced or solved, facts or explanations to be stated, and options to be selected for multiple choice or other objective items. Because a student may conceivably get a set of items either all correct or all incorrect, both perfect and zero scores are possible. The second method, backward counting, occurs in several specialized school subjects such as typewriting. Marks are subtracted from the perfect score by explicit rules according to the number of errors, mistakes, or faults.

In the third method, the teacher first makes a judgment about the quality of student work. The marks follows in time, and represents in degree, the qualitative judgment. Typically the teacher reads through, say, an essay, mentally locates its position on a quality continuum, and then **assigns** a number such as 17 by mapping this position on to the number line (between, say 0 and 20). The mental processes go on in private, and there are no "things" that are literally counted up to reach the total of 17. Teachers who are known for giving low scores are called tough or hard markers, other may be called fair or easy markers.

## Utility of Marks

Besides the familiarity which springs from a long tradition of use, numbers have a convenience and utility which goes a long way towards explaining their almost universal use. **Any** set of symbols, or course, simplifies recording. The symbols may be literal (A, B, C, ...), word phrases (Excellent, Very Good, ...), or numerical (7, 6, 5, ...). The present discussion is about marks as measures of academic achievement rather than about grade labels, but all symbol sets are designed to make it unnecessary for teachers either to retain samples of student work or to write detailed description records.

Mark scales can be constructed to cover any convenient range (such as 1-7, 0-20, or 0-100). The fundamental mathematical operations (addition, subtraction, multiplication, and division) may be readily performed, making it possible to (a) accumulate marks over a period of time, (b) combine marks from different sources such as tests, examinations, assignments, projects, and field study or laboratory write-ups, (c) calculate average scores for individual students or classes, (d) project a set of scores from one scale on to another for the purpose of making marks from different sources comparable, (e) scale a set of class scores to adjust for such factors as test difficulty, and (f) carry out research into the validity, reliability, and other aspects of scoring.

Apart from convenience, there are certain social concomitants to the use of marks, important ones being the objectivity and precision marks are commonly held to possess. In the popular mind, marks are accorded much the same status as physical measurements. Grading decisions based on the addition of sets of marks and predetermined cutoffs are generally accepted without question. In the rare event that a teacher's grade are challenged by a parent, it is usually the arithmetic that is questioned rather than the individual marks themselves or the arbitrariness of the grade cutoffs. Like the printed work, a mark is considered to have a natural integrity. Whether it deserves to, of course, is quite another matter. The fact remains that marks are associated with precision, and foster a sense of security.

It may be supposed that an assessment system based on marks would therefore be able to withstand a full legal challenge better than a global judgment against specified criteria and standards, because the marks are considered to be valid, reliable, and objective. But empirical evidence reported in the measurement literature for nearly a century has consistently shown that such confidence is often misplaced. In terms of criteria-based assessment, the issue is really whether a series of small-scale judgments, each of which is assigned a mark for later mechanical combination, is superior to a large-scale (grading) judgment made within a framework of specified criteria and standards (and so not made intuitively or against vague and variable in-the-head standards). Ultimately, the primary consideration is the quality of teachers' judgments, and the extent to which they give rise to accurate and reproducible grades. How the assessments are recorded is secondary.

## Underlying Assumptions

On the surface, it may appear that it is a simple matter to map achievement on to a number line and then use the standard mathematical operations. This process, however, makes at least four assumptions about both the mapping and the mathematical operations. The first three follow directly from the routine practice of adding marks together, taking the numbers at face value.

*Equivalence of units*. Where marks are obtained from different sources (laboratory work, field studies, tests, assignments, projects), an assumption is made that the marks represent equivalent units of achievement. A marks is a mark is a mark. A score of 15 out of 20 on an objective test is taken to be equivalent to 15 out of 20 for an essay.

*Interval scale property.* For a single instrument, it is tacitly assumed that marks are of uniform "worth" or "value" for all sections of the scale. This implies that an improvement in performance of 15 points from, say, 5 to 20 on a 100-point scale represents the same increment *in achievement* as an improvement from 85 to 100 points. In other words, it is assumed that the marks awarded are strictly proportional to the educational achievement across the whole range of the scale. When this assumption holds, the scale is said to possess *interval* properties. Alternatively, the mapping is said to be *linear.*

Two examples show how easy it is for this assumption to be violated. Consider first a multiple choice test of 100 items, each item having four options. The probability of getting a particular item correct by chance is obviously 0.25, the expected chance score on the test being 25. The educational significance or worth of a score between 0 and 25 is obviously much less than that of a score in the internal 75 to 100. If the measure of actual achievement is termed the *signal* and the effects of chance or guessing *noise*, another way of describing the lack of interval property is to say that the signal-to-noise ratio is low in the lower portion of the scale and high in the upper portion.

For the second example, consider essays marked on a 20-point scale. Many teachers routinely give work that is below par a mark of 8 or 9, provided the essay satisfies some basic minimum requirements such as being written in English (or some approximation to it) and being at least vaguely connected with the topic (as indicated by the essay writer's use of one or two key words from the task specifications). So a student "earns" say 8 marks essentially for trying. The scale of marks is effectively reduced to between 8 and 20, although even then some teachers never give 20 because nothing is, in their opinion, so perfect that it could not be improved. It is obvious that the assumption of an interval scale, in particular that the range 0 to 8 is comparable with the range 12 to 20, is a dubious one.

*Weightings*. The third assumption is that the importance of results from different instruments or dimensions of performance can be handled satisfactorily by a combination of mark allocation and weighting by instrument, and that it is appropriate to use uniform weightings for all levels of performance on the different instruments.

Actually, what is meant by weighting is not as simple as may at first appear. Suppose that a class of students sits for a practical test marked out of 25 and an examination marked out of 50. If the two components are to be weighted equally, it is commonly assumed that the marks for the practical test should be doubled before being added to the examination scores. Because both sets are then out of 50, so the thinking goes, the components are equally weighted. It is obvious, however, that the sets of scores are derived from different scales whose units are not necessarily commensurate and which in any case lack the interval property. Although the arithmetic is easy to do, it is less easy to describe what the equalization of maximum scores achieves, or what weighting itself means.

More sophisticated teachers do not fall into the trap of simply adjusting to get the required ratios among the maximum possible scores. Measurement textbooks point out that, when sets of scores which are not perfectly correlated with one another (the usual case) are added, one component may dominate the aggregate or composite score in the sense that the correlation between that component and the aggregate is much higher than with any of the other components. Another way of saying this is that the rankings based on the aggregate almost exactly correspond with the rankings on the dominant component, with the others seemingly having only minor influence.

Statistical analysis shows that, in general, the component which has the greatest spread of scores (as measured, say, by the standard deviation) has the greatest influence on the aggregate and so has an inbuilt heavier weighting. The advice given in most measurement textbooks is to multiply the several components by whatever factors are necessary to make the standard deviations match the desired weightings. For example, in the two-component case mentioned above, suppose that the standard deviations of the sets of raw scores were each about 5. The advice would be to add the raw scores as they are, ignoring the fact that one maximum score is twice the other. This would ensure that the resulting aggregate score ranks students in a way which is a reasonable compromise between the rankings on the individual components. But observe that in doing this, the concept of weighting takes on a special meaning, namely, that the two components have about equal influence on the aggregate *in the way they discriminate among students*. It is clear that both the interpretation of weighting, and the consequences of particular policies for weighting, are important matters in using marks and marking systems.

*Full compensation*. The final assumption has to do with the aggregation rule itself. Invariably, the rule is additive, involving summation of raw, scaled, or weighted scores. Addition is fully compensatory in the sense that a poor performance in any one area can be offset by correspondingly better performances in other areas. Even when only two assessment instruments are used, it is obvious that a given total (say 100) can be obtained in many different ways, because

$$100 = (90 + 10) = (50 + 50) = (10 + 90) = \ldots$$

The use of the additive rule as the principal basis for grading decisions implies that all possible combinations of marks which give rise to the same total represent equivalent achievements and are therefore equally valuable educationally.

## Marks and Criteria-based Assessment

Criteria-based assessment is geared towards comparing each student's work with fixed standards, not directly with the work of other students, and calls for a major reconceptualization of the assessment process. This reconceptualization is probably easier to achieve in subjects where teachers' qualitative judgments form a normal and natural part of assessment, such as in English, where the assignment of a score is made after a decision about quality. In such subjects, marks have an artificiality about them, a matter which leaves some teachers feeling uncomfortable about their use. For other subjects, in which the score is arrived at by counting, teachers may find it more difficult to disengage themselves from traditional practices and immerse themselves in judgments of directly qualitative kind. But having examined the assumptions underlying marks and marking systems, it is appropriate at this point to examine the role numbers should have in criteria-based assessment. Is it possible, or desirable, to use the traditional system of marks to record and report student achievement?

The answer is given here from several perspectives. It is shown in the discussion above that the technical assumptions underlying marks and marking systems are substantial, and as a group are not easily satisfied for any school-based assessment system, criteria-based or otherwise. There are, therefore, sound reasons for entertaining some reservations about the use of numbers in assessment generally, although these may not be serious enough to outlaw marks altogether. But the use of marks for recording and reporting is in many respects inimical to the aims of criteria-based assessment. Two practices in particular are logically incompatible with it in principle.

The first is the use of standard scores, which are defined by the mathematical transformation $\underline{z} = (\underline{X} - \underline{M})/\underline{S}$, where $\underline{z}$ is the required standard score, $\underline{X}$ is the raw score, $\underline{M}$ the group mean (average), and $\underline{S}$ the group standard deviation. Such standardization is often considered necessary in adjusting sets of scores for, say, test or task difficulty, in order to report achievement on a uniform scale, or to align score sets prior to forming weighted aggregates. There are three objections to this practice of standardization. Firstly, because calculating a standard score makes use of the mean and standard deviation obtained from the whole group of students, it clearly involves norm-referencing, and therefore has no place in assessment against fixed standards. Secondly, if standardization is used as part of the process of weighting components according to their standard deviations, the weighting is defined implicitly in terms of the different abilities of the components to discriminate among students, and is again totally irrelevant to the concept of criteria-based assessment. Finally, standardization of scores makes it unnecessary for the teacher to make the difficult analytic decisions about the performance of students *in relation to the specifications of the task and other aspects of the context*. Yet it is precisely these matters which should characterize a teacher's ability, as a professional, to make judgments about the achievements of students. Continued use of standard scores thus obviates the need for teachers to develop expertise in an important aspect of their work.

The second practice has to do with reporting student achievement to parents or prospective employers. It is well known that a raw score in isolation possesses little intrinsic meaning. For instance, to say that a student has received 73% in a test does not tell whether the level of performance is high or low. Among other things, the test may have been exceptionally easy, or the teacher may be given to awarding high marks for mediocre work. It is necessary, therefore, to provide some framework so that the 73% can be interpreted. The most common method of providing such a framework is to quote the class mean for the test (or place in class, or both) on school reports. This enables the performance of the student to be seen in relation to the performances of other students in the school. Again, this is clearly norm-referencing.

Other objections to marks are not so much incompatible in principle as likely to distort the practice of criteria-based assessment. Marks and scores are rarely considered to be neutral category labels. They immediately imply *measures* of achievement, and simply beg for addition and averaging. Teachers who have been brought up on marks and have used them uncritically for years may feel drawn to allocating marks according to the criteria and standards stated in the syllabus, and then using these marks simply to facilitate combination across different criteria. There are several disadvantages even in doing this.

(a) Once recorded, the criteria and standard from which the marks are derived recede into the background, and a "credit exchange" economy is established in which students exchange pieces of work for credits, which are then banked progressively. At the end of the course, the accumulated credits are withdrawn and the final course grade decided, without the necessity for a final reference to the criteria and standards for the course as a whole. This may produce two negative side effects. The first is that in most subject, some elements of a course are subsumed by later or more advanced elements or concepts. If the marks from earlier elements are added to those from later elements, some of the work is tested more than once, first explicitly, then implicitly as a component of another part, so that double counting occurs. Secondly, students have little motivation to develop evaluative experience themselves, and the formative potential of criteria-based assessment cannot by fully realized. Numbers provide such a highly efficient method of recording that neither teachers nor students may see any need to reconceptualize (i) the process of assessment in terms of comparisons against fixed standards, or (ii) the process of learning in terms of feedback and remediation.

(b) Marks encourage a quantitative notion of grading (or of pass-fail, often with 50 per cent as the pass score). Such quantitative thinking may not always take into account the *quality* of the work done. Indeed, in some classroom situations a greater quantity of work produces higher mark totals and hence higher grades. Where courses are structured as core-plus-electives, the speed with which some students may complete the core work at a satisfactory (or mastery) level may give them access to elective material. The final grade may then be dependent more on the scope of work attempted than on its quality in any absolute sense.

(c) Marking systems encourage inter-student comparisons (because it is so simple to compare marks or indeed symbolic grades on any scale), rather than comparisons of student achievements with the specified criteria and standards. Students may become interested only in the marks they obtain rather than in the reasons for those marks. Many students, in fact, claim only to look at the marks so that they can compare their performance against other students or against their previous marks, and not even to read teachers' comments.

(d) Marking systems encourage the use of pre-specified numerical cutoffs for grading decisions. The sum of component scores constitutes the grading decision, and conversely the judgments are validated entirely by reference to the summed scores and the cutoffs. Where only total score is used, it may be impossible to set supplementary minimum requirements for some aspects of a course because of the compensatory properties of the addition rule. It is well known that breaking down a difficult and complex judgment into a series of component judgments makes the assessor's task less difficult. However, it is necessary in criteria-based assessment to carry out a global retrospective review of a student's performance, judging the configuration of achievements on all relevant dimensions against the criteria and standards specifications. It has yet to be demonstrated whether a mathematical system can be developed which faithfully maps (i) the global achievement, (ii) allowable trade-offs on different criteria, and (iii) non-negotiable minima on some or all criteria, on to a final achievement scale which satisfactorily models teachers' carefully considered global judgments. There are reasons for believing that it is impossible in principle with any form of additive composition rule.

(e) When two or more sets of imperfectly correlated sub-scores are added, the totals tend to cluster about the middle. This phenomenon can be explained, at least, in part, from statistical theory and gives rise (with what seems inevitable regularity) to the familiar bell-shaped (or *normal*) distribution of scores. Whether the shape of the frequency distribution mirrors accurately the achievements among students, or is in fact an artifact of the additive aggregation rule, then goes unexamined. This in turn may reinforce the expectation that grades should similarly follow a normal distribution.

## Conclusion

Marks and marking systems are in almost universal use in secondary schools. This is due partly to tradition and familiarity with numbers, partly to convenience, and partly to the absence of workable alternatives. An analysis of the underlying assumptions shows that numerical marking systems enjoy a status that is higher than they strictly deserve. The use of marks in criteria-based assessment is inappropriate for two sets of reasons. Firstly, the assumptions are not generally satisfied in any form of school-based assessment, and secondly, the use of marks as currency in grade-exchange transactions diverts attention away from criteria, standards, and the processes of qualitative appraisals, and to that extent is educationally counterproductive.