Guidelines for workplace learning from the University of Tasmania: www.utas.edu.au/tl/supporting/workplace_learning.html

On lifelong learning

Knapper, C. and Cropley, A. (2000) *Lifelong Learning in Higher Education*. London: Kogan Page.
http://www.lifelonglearning.co.uk/
http://www.adelaide.edu.au/clpd/materia/leap/leapinto/LifelongLearning.pdf

Where do we start selecting? Knapper and Cropley's book is one of the classics in this area. The two home pages are of lifelong learning sites, one in the UK, the other in Australia, with plenty of links.

On problem-based learning

Boud, D. and Feletti, G. (eds) (1997) *The Challenge of Problem-based Learning*. London: Kogan Page.
Savin-Baden, M. (2000) *Problem-based Learning in Higher Education: Untold Stories*. Buckingham: The Society for Research into Higher Education/Open University Press.
Research and Development in Problem Based Learning. The Australian Problem-Based Learning Network c/o PROBLARC, CALT, The University of Newcastle, NSW 2308.

Boud and Feletti contains contributions by users in many different areas. Savin-Baden introduces a little-discussed aspect: what happens *inside* when teachers and students experience PBL. Both books are important for anyone seriously interested in PBL. The last is a serial publication of the Australian Problem-Based Learning Network, which holds biennial conferences, of which these volumes are the proceedings.

Waters, L. and Johnston, C. (2004) Web-delivered, problem-based learning in organisation behaviour : a new form of CAOS, *Higher Education Research and Development*, 23, 4: 413–431.

An e-version of PBL in teaching organizational behaviour is based on *Case Analysis of Organisational Situations*.

PBL in biology (20 case examples): www.saltspring.com/capewest/pbl.htm
PBL in physics, chemistry, biology and criminal justice: www.udel.edu/pbl/problems
PBL in engineering: http://fie.engrng.pitt.edu/fie2001/papers/1102.pdf

9

Aligning assessment tasks with intended learning outcomes: Principles

What and how students learn depends to a major extent on how they think they will be assessed. Assessment practices must send the right signals to students about what they should be learning and how they should be learning it. Current practice, however, is distorted because two quite different models of summative assessment have, for historical reasons, been confused and the wrong signals to students are often sent. In this chapter, these issues are clarified. We examine the purposes of assessment, the relation between assessment and the assumed nature of what is being assessed, assessing for desirable but unintended or unexpected learning outcomes and who might usefully be involved in the assessing process. The underlying principle is that the assessment tasks should comprise an authentic representation of the course ILOs.

Formative and summative assessment

There are many reasons for assessing students: selecting students, controlling or motivating students (the existence of assessment keeps class attendance high and set references read), satisfying public expectations as to standards and accountability, but the two most outstanding reasons are for *formative feedback* and for *summative grading*. Usually – and perhaps unfortunately – both are referred to as types of 'assessment'. Both are based on seeing how well students are doing or have recently done, which is what assessment is, but the purposes of the two forms of assessment are so different.

In formative assessment, the results are used for *feedback* during learning. Students and teachers both need to know how learning is proceeding. Formative feedback may operate both to improve the learning of individual students and to improve the teaching itself. Formative feedback is inseparable from teaching: as we have already noted (p. 97), the effectiveness of different teaching methods is directly related to their ability to provide formative feedback. The lecture itself provides little. The improvements to the lecture

mentioned in Chapter 7 were almost all formative in function: they got the students learning actively and feedback was provided on their activity, either from teacher or from peers. Formative feedback is a powerful TLA that uses error detection as the basis for error correction: if error is to be corrected, it must first be detected. Thus, students must feel absolutely free to admit error and seek to have it corrected. Students also need to learn to take over the formative role for themselves, just as writers need to spot error and correct it when editing a text by reflecting critically on their own writing. Self- and peer-assessment are particularly helpful TLAs for training students to reflect on the quality of their own work.

In summative assessment, the results are used to grade students at the end of a course or to accredit at the end of a programme. Summative assessment is carried out after the teaching episode has concluded. Its purpose is to see how well students have learned what they were supposed to have learned. That result, the grade, is final. Students fear this outcome; futures hinge on it. They will be singularly unwilling to admit their mistakes. Error no longer is there to instruct, as in formative assessment; error now signals punishment. This difference between formative and summative reminds us that continuous assessment (see later) is problematic when it is used for both formative and summative purposes. What then does the student do about admitting error? This is one area where the same word 'assessment' leads to confusion.

Nevertheless, there is one similarity: in both we match performance as it is, with performance as it should be. When the student is aware of the immediate purpose to which it is being put, the same task can act as a TLA, in the formative sense, and as the assessment task when it is time to do the summative assessment: 'When the chef tastes the sauce it is formative assessment; when the customer tastes, it is summative' (Anon.). Figure 9.1 places tasting the sauce in a classroom context.

Say four topics are to be learned in a semester. The ILOs of each are symbolized as IL01, IL02, IL03 and IL04. At the start of the semester (labelled 'baseline') students enter with little or some knowledge, which the TLAs nurture until the end of the semester. Formative assessment checks that growth and sees that it is on track. Then it is time to see where each student now stands with respect to each of the four topics; this is the task of summative assessment. Finally, there is the administrative matter of converting those four positions into a grade, taken here as A, B, C and D.

A caution in interpreting Figure 9.1. While the same assessment task may be used formatively throughout the course and summatively at the end, it must be clear to the students when it is being used for what purpose. To use it for *both* formative and summative purposes, as may happen in continuous assessment, creates a conflicting situation for the students: they are being asked to display and to hide error simultaneously. When assessment is continuously carried out throughout a course, and it is intended to use some of the results summatively, the students must be told *which* assessment events are formative and which summative. They can then decide how they will handle the task to best advantage.

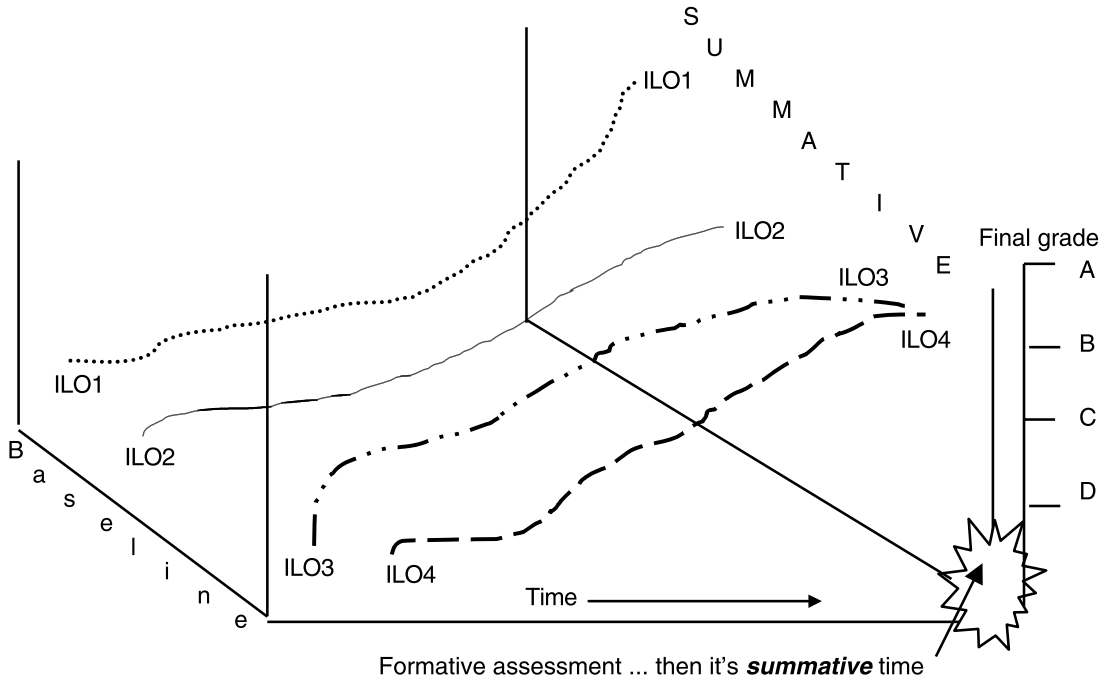

Figure 9.1 Learning in four topics and their formative and summative assessment

Two quite different models of summative assessment have become confused. Our intended learning outcome of this chapter is that readers will unconfuse themselves so that they may apply the appropriate model to their own assessment practices.

But before we do this, Task 9.1 present six dilemmas ('cases') relating to assessment practice. Go through these, writing your responses down. When you have completed this chapter you might like to revisit what you wrote to see if your thoughts might have changed.

Task 9.1 Some cats to place among your collegial pigeons: six assessment dilemmas for you to consider

Case 1. Misunderstanding the question

You are assessing assignments and find that one student has clearly misunderstood the question, the only one to have done so. It is now past the due date for handing in. If you assess it as it is, she will fail. What do you do?

a Fail her.
b Hand it back, explain that she has misunderstood and give her an extension.
c As in (b), but assess it pass/fail only or deduct a grade.
d Set her another assignment, to be assessed later. Meantime record 'result withheld'.
e Other. What?

What are the reasons for your decision? _____

_____

_____

_____

Would you have decided differently if she would otherwise graduate with distinction?

_____

_____

Case 2. Grading on the curve

The guidelines for awarding a grade of A are outlined in a programme document:

> *Outstanding*. Demonstrates thorough understanding and interpretation of topics and underlying theories being discussed, and shows a high level of critical thinking and synthesis. Presents an original and thorough discussion. Well organized and structured, fluently written and correctly documented. There is evidence of substantial studies of the literature.

You use these guidelines in grading the assessment tasks of your class of 100 students and find to your delight that 35 (35%) meet these criteria, so you award A to all of them. Your departmental head, however, is unhappy about this because you are 'not showing enough discrimination between students and we don't want this department to get a reputation for easy marking'. The results have not been announced yet, so he suggests that you regrade so that only 15% of your students are given an A. What do you do? Why?

a You agree you must have been too lenient, so you do as he says, giving A to the top 15 only, the remaining of the original As being given B.
b You compromise, splitting the difference: you give As to 25 students.
c You say something like: 'Sorry, but the guidelines are clear. I must in all conscience stick with the original. The conclusion to be drawn is that this was an exceptionally good group of students and that they were taught well.'
d 'I must stick with the guidelines. However, I am prepared to entertain a second opinion. If I can be persuaded that I have been too lenient, I will change my grades.'
e Other.

_____

_____

_____

_____

Case 3. A matter of length

It is policy that the maximum word length of assignments is 1000 per credit point. You are teaching a 2-credit point module. One of your better students has handed in an assignment of 2800 words. What do you do and why?

a Count up to 2000 words, draw a line and mark or assess up to that point only.
b Hand it back to the student with the instructions to rewrite, within the limit, with no penalty.
c As for (b) but with a penalty. (What would you suggest?)
d Hand it back unassessed.
e Assess or mark it but deduct a grade or part-grade, or marks, according to the excess.
f Other.

_____

_____

Would your decision have been any different if it were a poor student?

_____

_____

Case 4. Exam strategy

You are discussing the forthcoming final exam with your first year class. You explain that, as usual, there will be five sections in the paper, each section covering an aspect of the course, and there are two questions per section. They are to choose one of the two, making a total of five questions, to be completed in three hours. You alone will be doing the assessing. A student asks: 'If I think I will run out of time, is it better to answer four questions as best as I can, or to attempt all five, knowing I won't finish most questions?'

What do you say in reply and why?

_____

_____

_____

Case 5. Interfering with internal affairs?

You are the head of a department that has decided to use problem-based learning in the senior level subjects. In PBL, the emphasis is on students applying knowledge to problems, rather than carrying out detailed analyses of the research literature, as has been the tradition in the past. Faculty regulations require you to set a final examination for

the major assessment of the course, despite your own judgment and that of your staff that this format is unsuitable for PBL. It is therefore decided that the final exam will contain questions that address application to problem solving rather than questions that require students to demonstrate their familiarity with the literature.

On seeing the paper, however, the external examiner insists that the questions be reworded to address the research literature. You argue, but he insists that 'academic standards' must be upheld. If they are not reworded, you know that he will submit an adverse report to the academic board, where there are vocal critics of your foray into PBL.

What do you do?

_____

_____

_____

_____

Case 6. What is the true estimate of student learning?

A department is trying to arrive at a policy on the proportion of final examination to coursework assignments. In discussing the issue, the head collates data over the past few years and it becomes very clear that coursework assessments are consistently higher than examination results. In discussing this phenomenon, the following opinions are voiced. Which argument would you support?

a Such results show that coursework assessments may be too lenient and because the conditions under which they are undertaken are not standardized, and are unsupervised, the results may well be inflated by collaboration and outright plagiarism. Examination conditions control for these factors. Therefore final exams must be a higher proportion of the final grade than coursework assessments.

b The conditions under which final examinations are conducted are artificial: working under time pressure, little and often no access to tools or data sources, and mode of assessment limited to written expression or MCQ, means that exam performances are sampling only a narrow range of students' learning. Therefore coursework assessments must be a higher proportion of final grade than exams.

c Other. What?

_____

_____

_____

Effects of assessment on learning: Backwash

We teachers might see the intended learning outcomes as the central pillar in an aligned teaching system, but our students see otherwise: 'From our students' point of view, assessment always defines the actual curriculum' (Ramsden 1992: 187). Students learn what they *think* they will be tested on. This is *backwash*, a term coined by Lewis Elton (1987: 92), to refer to the effects assessment has on student learning, to the extent that assessment may determine what and how students learn more than the curriculum does.

Backwash is almost invariably seen negatively (Crooks 1988; Frederiksen and Collins 1989). Recall the 'forms of understanding' that Entwistle and Entwistle's (1997) students constructed to meet presumed assessment requirements (see pp. 74–5). Negative backwash always occurs in an exam-dominated system. Strategy becomes more important than substance. Teachers actually teach exam-taking strategies, such as telling students to attempt all questions even if they don't finish any because they gain more marks than by thinking deeply over a question and providing a complete answer. Students go through previous papers, best-guessing what questions they will encounter and then rote learning answers to them. This sort of backwash leads inevitably to surface learning. Yet learning for the assessment is also inevitable; students would be foolish if they didn't. So, what do we do about it?

In fact, backwash can work positively, encouraging appropriate learning. This is when the assessment is aligned to what students should be learning (Figure 9.2).

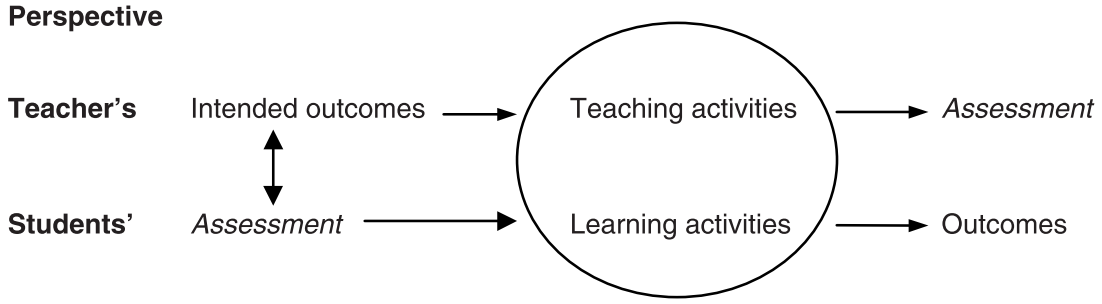

Figure 9.2 Teacher's and student's perspectives on assessment

To the teacher, summative assessment is at the end of the teaching–learning sequence of events, but to the student it is at the beginning. If the intended outcomes are reflected in the assessment, as indicated by the downward arrow, the teaching activities of the teacher and the learning activities of the student are both directed towards the same goal. In preparing for the assessments, students will be learning the intended outcomes.

It sounds easy, but there is a long tradition of thinking about assessment, and some time-honoured assessment practices, that complicate matters. In this chapter, we clarify some of the conceptual issues involved; in the next, we deal with designing and grading assessment tasks for declarative knowledge

and, in the chapter after that, designing and grading assessment tasks for functioning knowledge.

Measurement model of assessment

Two quite different models of assessment underlie current thinking and practice: the *measurement* model and the *standards* model (Taylor 1994). Understanding the difference between the two models is basic to effective assessment.

In the Chinese Han Dynasty in the 4th century BC, the purpose of education was selective. Students were required to master a huge classical curriculum, in order to put into effect Confucius' belief that 'those who excel in their study should become officials' (quoted in Zeng 1999: 21). The winners, however lowly their background, were motivated by a rich prize: a lifetime of wealth and prestige. The idea was to select the best individuals in terms of stable characteristics: 'not only intelligence, but also character, determination, and the will to succeed' (Zeng 1999: iv).

Twenty-three centuries later, psychologists in the 19th century also became interested in sorting people out. Sir Francis Galton (1889) found that physical and mental differences such as height, weight and performance on various mental tests, which he called 'traits', were distributed in 'an unsuspected and most beautiful form of regularity' (Galton 1889: 66). He was of course referring to the normal curve, a distribution that occurs inter alia as a result of the extent of the polygenetic inheritance of such traits. Galton's assumptions, not only about statistical techniques, but also about the inheritance of ability and of educability, were built into the burgeoning industry of mental testing in the early part of the 20th century.

Educability was assumed to be about how bright people were and, back to the Han Dynasty, education was seen as a device for sorting people out: usually the brightest, but sometimes to sort out those who weren't educable in normal schools. The present so-called 'parametric' statistical procedures such as correlation and factor analysis were based on Galton's work and are used for constructing educational tests, establishing their reliability and validity and interpreting test scores. Taylor (1994) refers to this individual differences model as 'the measurement model' of educational assessment.

The measurement model was originally designed by psychologists to measure stable traits and abilities and to express that measurement along a graduated scale so that individuals could be compared, either against each other or against population norms. This is fine for research, or for diagnosis when dealing with individuals – for example to say how atypical a person is on reading ability – but the model was hijacked and applied to assessing *educational* outcomes in the form of norm-referenced assessment (NRA).

In NRA, results of assessment are reported in terms of comparisons between students. The rank order is the simplest example, which tells who

performs better than who, but there are sophisticated versions of NRA, such as grading on the curve, which we discuss later.

For now, let us examine some of the assumptions the measurement model is based on when applied to assessing classroom learning.

Some assumptions of the measurement model

(*Note*: Beware the following subheadings: all are either wrong or misleading.)

Knowledge can be quantified

Measurement requires that the learning outcomes of individual students are quantified as scores along a single dimension or continuum so that individuals may be compared with each other. In practice, this means that learning is evaluated according to *how much* material has been learned correctly. Good learners know more than poor learners. The Level 1 view of teaching makes essentially quantitative assumptions, as we noted in Chapter 2: teaching involves transmitting the main points, assessment involves marking students on their ability to report them back accurately. The uni- and multistructural levels of the SOLO taxonomy are quantitative, where learning is a matter of finding out more and more about a topic. But if you assess only using quantitative techniques, what happens to the higher levels of the SOLO taxonomy: to our ILOs addressing critical analysis or hypothesizing?

Percentages are a universal currency

One of the commonest forms of quantification is the percentage, derived either as the ratio of number right to maximum possible multiplied by 100, or as sets of ratings the maxima of which total 100. When this transformation is carried out, it is assumed that percentages are a universal currency, equivalent across subject areas and across students, so that different students' performances in different subjects can be summed, averaged and directly compared. This is completely unsustainable, yet that doesn't stop university senates having long and earnest debates about one faculty using 75% as the cut-off for an A grade and another using 70% as the cut-off: 'We must level the playing field across faculties! It's not fair if it's easier to get an A in arts than it is in science!' Such debates are silly: they are trying to extract certainty from the unknowable. There is simply no way of knowing if 75% in physics is the 'same standard' as 75% in history; or even if a student's result of 75% in Psychology 201 this year represents an improvement over 70% the same student obtained in Psychology 101 the previous year.

Educational tests should be designed to clearly separate the high and low scorers

Measurement experts used to maintain that a good attainment test yields 'a good spread', following the bell curve (back to Galton). However, grades follow the bell curve only if two conditions apply: that ability is normally

distributed, and that ability is the sole determinant of academic attainment. But the ability of our students is not likely to be normally distributed because students are not randomly selected – not quite yet, anyway. And neither is ability the sole determinant of students' learning outcomes. Other factors are called 'teaching' and 'learning'. As argued in Chapter 1, good teaching narrows the initial gap between Robert and Susan therefore producing a *smaller* spread of final grades than that predicted by the initial spread of ability. The distribution of results after good teaching should not be bell shaped but skewed, with high scores more frequent than low scores. At university level there is therefore every reason *not* to expect a bell curve distribution of assessment results in our classes. Forcing assessment results to follow the curve actually prevents us from seeing how students are really performing.

Quantitative approaches to assessment are scientific, precise and objective

Numbers mislead. The measurement model yields an extended continuous scale that invites us to make minute distinctions between students, but we have to be careful. The error of measurement in our usual class sizes is bound to be rather more than one percentage point. Worse, the way we use the scales prevents them from being equal interval scales, where the difference between any two adjacent numbers is the same as any other two. This is an *essential* property if we are to average and accumulate marks. The difference between 73 and 74, say, must be the same as the difference between 79 and 80, if marks are to be added or averaged. But the difference between 79 and 80 often becomes zero if first class honours is awarded to a dissertation of 79 marks when the cut-off is 80 (see Box 9.3, p. 182). Many times, teachers and boards of examiners are faced with the borderline case and argue that as the scale is not accurate to one mark, we'll give the student the benefit of the doubt. This, however, makes our scale elastic, distinctly more rubbery at some points along the scale than at others.

Do such decisions show how human we are or just how sloppy? We are both and neither. We are being wonderfully inappropriate, like cooking dinner in the chemistry lab. The precision of the parametric measurement model is just as out of place in the classroom as is weighing sugar in milligrams. It is worse, actually, because the procedure of quantifying qualitative data, such as shifts in students' understandings, requires arbitrary judgments as to what is a 'unit', what is 'worth' one mark, what is worth five or however many marks. These judgments are not only subjective; they often do not even have an explicit and examinable rationale, beyond a vague norm referencing: 'I am marking out of five, this is the best so it gets five, this is average so it gets three marks.' What the *criteria* are that allow the judgement that this one is 'best' and that one 'average' may not be examined.

What happens, then, is that a series of independent minor subjective judgments – a mark for this, a mark for that – accumulate. The big decision – pass or fail?, first class or upper second? – is made on the aggregate of numbers, which includes the aggregate of error in all those minor judgments. That big decision should be made, not on the accumulation of

unknowably flawed minor judgments, but on a reasoned and publicly sustainable judgment about the performance itself. This requires a holistic judgment made on publicly stated criteria.

The application of a precise, scientific model to an area where it does not apply cannot be scientific.

University education is selective

Comparing students with each other assumes that universities are a selective device to find the intellectuals in the population, as in Han Dynasty China, or that the purpose of the undergraduate years is to weed out the 'pass' level students from the potential postgraduate research students.

The only place for assessing students selectively in the university context is for entry to university or to graduate school. At entry, a convenient estimate of scholastic ability is obtained by summing a student's best three, or best five, HSC or A level subjects, with or without adjustments for second attempt. What you get is a measure of scholastic ability, which is robust enough to allow direct comparisons between students in different subject areas. It is rough, but it works over large numbers. Once students have been selected, however, the aim of undergraduate teaching is to get students to learn what is in the curriculum, an enterprise in which the measurement model has no place.

But shouldn't the entry into university, and especially into graduate school, be based on whether the students are able to meet the criteria or standards necessary for doing graduate work? You don't answer that question by comparing students with one another.

The above assumptions give rise to some common practices.

Grading on the curve

After ranking, a common form of norm-referenced assessment is 'grading on the curve'. The top 10% of the class, say, are awarded 'high distinction', the next 15% 'distinction', the next 25% 'credit' and 45% 'pass'. The results will appear to be stable from year to year and from department to department. If there is a query from the odd student about the grade awarded, it is easy to point to an unarguable figure: all objective, very precise. 'You didn't earn enough marks to beat the others. They were too good for you. Sorry.'

The very term 'high distinction' is comparative, applicable only to that blessed few who are highly distinguished. This puts the brake on the number of HDs awarded. Even if one-third of the class met the criteria set for obtaining a high distinction, it would be seen by colleagues on the board of examiners, with the bell curve tolling in their heads, as a contemptible fall in standards, not as it should be a cause for congratulation. Rather than calling the highest grade a 'high distinction', the neutral term 'A' makes it easier to accept that a high proportion of students could reach that high standard.

Many people, teachers, administrators, and even students, feel it 'fitting' that a few should do extremely well, most should do middling well and a

few do poorly, some failing. This feeling comes straight from the assumptions that ability determines learning outcomes and that ability is normally distributed. Both assumptions are untenable, as we have seen.

Unfortunately, grading on the curve is so easy. All you need is a test that will rank order the students – a quick and dirty MCQ will do – and then you simply award an A to the first 10%, B to the next 25%, or whatever has been decided, and so on. Alignment is irrelevant.

Grading on the curve also appeals to administrators, because it conveys the impression that standards over all departments are 'right', not too slack, not too stringent, so that a few do really well, most middling and a few poorly: we have got it *right*, year after year. But that result is an artefact: the distribution has been defined that way, whatever the actual results in any given year or department.

Grading on the curve precludes aligned teaching and criterion-referenced assessment. It is a procedure that cannot be justified on educational grounds.

Marking

Marking is an assessment procedure that comes directly from quantitative assumptions and is so widespread as to be universal. It is, however, a procedure that needs to be examined closely. Marking is quantifying learning performances, either by transforming them into units (a word, an idea, a point), or by allocating ratings or 'marks' on a subjective if not arbitrary basis. For marking to be acceptable, we have seen that one mark must be 'worth' the same as any other, so that they can be added and averaged and a grade is awarded on the number of marks accumulated. Two most peculiar phenomena are associated with marking:

1. Half the total number of marks available is almost universally accepted as the pass mark.
2. It does not matter *what* is correct, as long as there are enough of them.

Multiple-choice tests enact these assumptions exactly. Learning is represented as the total of all items correct. Students quickly see that the score is the important thing, not how it is comprised, and that the ideas contained in any one item are of the same value as in any other item. The strategy is to focus on the easy or trivial items; and of the alternatives you don't know, check the ones that seem vaguely familiar. You'll almost certainly get more than half correct – and by defintion you'll pass.

The essay format, technically open ended, does not preclude quantitative means of assessment. When multiple markers use marking schemes, they give a mark or two as each 'correct' or 'acceptable' point is made, possibly with bonus points for argument or style. This too sends misleading messages to students about the structure of knowledge and how to exploit its assessment. A good example is the strategy in timed examinations of attempting all

questions and finishing none. The reasoning is that the law of diminishing returns applies: the time spent on the first half of an essay nets more marks than the same time spent on the second half. The more facts the more marks, never mind the structure they make. But students don't learn 'marks', they learn such things as structures, concepts, theories, narratives, skills, performances of understanding. These are what should be assessed, not arbitrary quantifications of them. It is like examining architects on the number of bricks their designs use, never mind the structure, the function or the aesthetic appeal of the building itself.

Assessment separated from teaching

In the measurement model, assessment is a standalone activity, unrelated to teaching as such. Accordingly, it attracts its own context and culture. One feature is the need for standardized conditions including the same assessment tasks for all, a necessary condition when students are to be compared with each other. Guaranteeing standardized procedures leads to a Theory X, bureaucratic assessment climate: emphasis on decontextualized assessment tasks that address declarative, not functioning, knowledge.

In universities that work in this way, teaching occupies the greater part of the academic year, assessment a frantic couple of weeks at the end. Both the present writers can recall, now with shame, not even thinking about the final examination until the papers were due to be sent to the central examinations section. You teach as it comes, you set an examination, the examination centre invigilates it for you, you allocate the marks.

Alignment doesn't come into it.

Effects of backwash from the measurement model

Measurement model procedures send unfortunate messages to students:

- *The trees are more important than the wood.* Maximizing marks is the important thing, not seeing the overall structure of what is being learned. Put another way, the measurement model encourages multistructural thinking, not relational or extended abstract.
- *Verbatim responses will gain marks.* Although a verbatim replay of a unit in the text or in the lecture may not be very noble, it has to be given some credit when using a multistructural marking scheme, given cheating has been ruled out. This happens even when the teacher warns that verbatim responses will be penalized (Biggs 1973 (regrettably)).
- *Success or failure is due to factors that are beyond the student's control.* An individual's result under NRA depends on the competition, who is more able. Thus, in the event of a poor result, the student can either blame bad luck or, more damagingly, come to the conclusion that he or she is simply not as able as other people. Students can't do anything about luck or ability,

so why bother? The attribution under the standards model is different: 'Here is what I am supposed to have achieved, I didn't achieve it, so what went wrong?' The answer to that could be: 'I didn't put in enough effort', 'I didn't know how to do it' but, at worst: 'I am dumb.' The first two attributions are under the students' control and they can do something about doing better next time. The last couldn't be more discouraging.

The case against the measurement model is pretty convincing, so why do its procedures remain? Box 9.1 suggests some answers.

Box 9.1 Why measurement model procedures remain

1 *Tradition, habit.* Why question what has worked well in the past, especially when administrative structures and procedures make change difficult?

2 *Bureaucratic convenience*

- Dealing with numbers gives the illusion of precision. Any appeal or disagreement is over trivial issues. Let the numbers make the big decisions.
- Grading on the curve gives the illusion of constant standards, no egregious departments or results.
- The language of percentages is generally understood (another illusion).
- Given the tight security of exams, avoidance of plagiarism can be assured.
- Combining results from different departments needs a common framework: the percentage and normalized scores (both illusions, see earlier point).

3 *Teaching convenience*

- You teach, the exam questions can be left until well into the teaching, exams section will see to the details. It is flexible on coverage, what questions you set.
- You can easily average and combine marks across tasks and across courses.
- You can use marks for disciplinary purposes (deduct for late submission).
- It's easier to argue numbers with students in case of dispute than to argue 'subjective' structures.

4 *Genuine belief in the measurement model.* My job *is* to sort the sheep from the goats.

Let us now turn to the alternative, the standards model.

Standards model of assessment

The standards model of assessment is designed to assess changes in performance as a result of learning, for the purpose of seeing what, and how well, something has been learned. Such assessment is criterion-referenced (CRA), that is, the results of assessment are reported in terms of how well an individual meets the criteria of learning that have been set. This model is the relevant one for assessment at university (Taylor 1994). The point is not to identify *students* in terms of some characteristic, but to identify *performances* that tell us what has been learned, and how well. Unlike in NRA, one student's result is quite independent of any other student's.

In 1918, R.L. Thorndike made it very clear that CRA was most appropriate for educational purposes, and predicted that CRA would displace NRA from public schooling (Airasian and Madaus 1972). He was right about the first point, but, unfortunately, his prediction was wrong. The idea still lurks that education *is* a selective exercise, and that norm-referenced examinations are appropriate. But even where this idea is not explicit, the procedures of constructing and administering tests, establishing reliability and validity and interpreting and reporting test scores are based on parametric statistics, as if the biological assumptions of polygenetic inheritance, which produce the normal curve, are appropriate for educational assessment. As already argued, for purposes of classroom assessment such statistics as the correlation and the usual tests of reliability and validity are entirely inappropriate. Reliability and validity of assessments are important, but they have entirely different meanings in the standards model (pp. 188–90).

Outside educational institutions, the standards model is assumed whenever anyone teaches anyone else anything at all. The teacher has a standard, an intended outcome of their teaching, which the learner is to learn satisfactorily. Parents intend their children to learn to dress themselves to a given standard of acceptability, swimming instructors have standards they want their learners to achieve. Parents don't lecture a toddler on shoe tying, and give a multiple-choice test at the end to see if their child ties her shoes better than the kid next door. The parent's ILO, the teaching/learning activity and the assessment are all the same: it is tying a shoe. In the case of driving instruction it is driving a car. The alignment is perfect. Outcomes-based teaching and learning is placing this approach back into the institution.

The logic is stunningly obvious: Say what you want students to be able to do, teach them to do it and then see if they can, in fact, do it. There is a corollary: if they cannot do it, try again until they can. This principle is used in 'mastery learning' (Bloom et al. 1971) and the Keller Plan, a mastery model for universities (Keller 1968). Students are allowed as many tries at the assessment as they need – within reason – in order to pass the preset standard. Some students pass in short order, others take longer. The main objections to mastery-learning models were not to the principle, but to the fact that the preset criteria were defined quantitatively, mainly because quantitative criteria are easy to define. In one study with high school biology

students, the Roberts who focused on memorizing detail performed well in such a mastery-learning approach, but not the Susans who were bored stiff (Lai and Biggs 1994).

Such objections do not apply when the standards are defined *qualitatively*. Qualitative assessment does not directly address the question of *how much* the student knows, but *how well*. This requires an explicit classification of learning quality that needs to be derived for each topic or skill taught. The SOLO taxonomy is a general model of learning quality that can be adapted to suit particular content (see Chapter 5).

Let us now look at the assumptions needed to make the standards model of assessment work.

Some assumptions of the standards model

We can set standards (criteria) as intended learning outcomes of our teaching

Yes we can, as outlined in Chapter 5. If the intended learning outcomes are written appropriately, the job of the assessment is to state how well they have been met, the 'how well' being expressed not in 'marks' but in a hierarchy of levels, such as letter grades from 'A' to 'D', or as high distinction through credit to conditional pass, or whatever system of grading is used. Deciding at the level of a particular student performance is greatly facilitated by using explicit criteria or rubrics (examples on pp. 210–Table 10.2, 214–Table 10.4, 226–Table 11.2). These rubrics may address the task, or the ILO.

Different performances can reflect the same standards

While standardized conditions are required when individuals are to be compared to each other, when we are seeking to find the optimum performance of individuals, the more standardized the conditions the less valid the test is likely to be for any given individual. Individuals learn and perform optimally in different conditions and with different formats of assessment. Some work better under pressure, others need more time. As in professional work itself, there are often many ways of achieving a satisfactory outcome. Individual students demonstrate their best work in different ways; assessment tasks such as portfolios allow for that.

Teachers can judge performances against the criteria

This is critical when using the standards model but it is skirted when using the measurement model. In the latter, teachers need to answer the following question: 'How many marks do I give this section?' and in the former: 'How well does this performance as a whole meet the criteria for high distinction (or whatever)?' In order to make these holistic judgments teachers need to know what is poor quality performance, what is good quality and why.

Constructive alignment operates on these same assumptions and addresses how they may work in practice.

Norm- and criterion-referenced assessment: Let's get it straight

Differences between NRA and CRA

Because of the universality of many NRA practices in assessing students, and the educational logic of CRA, we should be clear about the differences. To recap briefly:

1. In NRA, the results are expressed in terms of comparisons between students after teaching is over. CRA results are expressed in terms of how well a given student's performance matches criteria that have been set in advance.
2. NRA makes judgments about *people*, CRA makes judgments about *performance*.

Task 9.2 presents a criterion-referenced test to sort the sheep from the goats (joke).

Task 9.2 NRA or CRA?

A teacher assesses two students in a CRA system and notes that Robert has been awarded a B and Susan an A. On a recheck of the papers, the teacher notes with a shock that Robert's paper *is* as good as Susan's! He is reassessed and given an A too.

Is this now NRA (comparing students) or CRA (judging on standards)? Why?

_____

_____

The answer is at the end of this chapter.

A summary of the differences between CRA and NRA is captured in Table 9.1, which lists a lexicon of NRA and CRA words. The only word common to both? Summative assessment.

Nevertheless, it is easy to blur the two models. Box 9.2 (p. 181) represents a valiant attempt by an arts faculty at one university to move towards the standards model. Previously, a marks system was used to define 'A+', 'A' and 'A–' and so on, and the attempt was made at faculty board to devise a scheme that defined the grading categories, avoiding marks. The following was issued to all teachers in the faculty.

You work out what the problem is. Then turn to Box 9.3 (but no peeking!) (p. 182).

Table 9.1 Two lexicons

Norm-referenced assessment
Mark, percentage, decile, rank order,* summative assessment, decontextualized assessment, standardization, 'fairness', quantitative, average, grade-point average, normal/bell curve, normal distribution, grading on the curve, a good spread of scores, parametric statistics, test–retest reliability, internal consistency, discrimination, selection, competition, high flier, ability
Criterion-referenced assessment
Assess, authentic/performance assessment, contextualized, standards, formative assessment,* summative assessment, criteria, individualization, optimal performance, student-centred, qualitative, grading categories, ILOs, alignment, judgment, distribution free, non-parametric statistics, effort, skill, learning, competence, expertise, mastery

* The one word in common!

A double problem

Despite the prevailing norm-referenced cast of mind at undergraduate level, the sheer logic of criterion-referenced assessment is generally seen in assessing theses and dissertations. We expect a dissertation to display certain characteristics: coverage of the literature, definition of a clear and original research question, mastery of research methods, and so on. The categories of honours (first class, upper second, lower second) originally suggested qualities that students' work should manifest: a first was qualitatively *different* from an upper second, it was not simply that the first got more sums right. Today, this approach might be in jeopardy, as these categories seem increasingly to be defined in terms of ranges of marks, which is unfortunate. In Box 9.4 (p. 183) we see a doubly unfortunate instance: defining the level of honours in terms of marks, and allowing non-academic factors to influence the judgment of academic quality.

In the standards model, and in constructive alignment in particular, this double problem could not occur. The ILOs would refer to academic qualities only, not sexual harassment, lateness or anything else, and the assessment would be aligned to those ILOs. There are other and more appropriate ways of dealing with the non-academic issues than by adjusting final grades.

Some important concepts in assessment

Authentic and performance assessment

In assessing functioning knowledge in particular, the assessment tasks need to represent the knowledge to be learned in a way that is authentic to real life. Verbal retelling is not often authentic; for example, we do not teach

Box 9.2 How Faculty Office suggests final grades should be determined (and the best of British luck!)

The following guidelines were issued to all staff in the faculty. They were to use these in arriving at their final grade distributions:

A (A+, A, A–)	Excellence, up to 10% of students. The student must show evidence of original thought as well as having a secure grasp of the topic from background reading and analysis
B (B+, B, B–)	Good to very good result, achieved by next 30% of students who are critical and analytical but not necessarily original in their thinking and who have a secure grasp of the topic from background reading and analysis Occasionally, a student who shows originality but is less secure might achieve this result
C (C+, C, C–)	Satisfactory to reasonably good result. The students have shown a reasonably secure grasp of their subject but probably most of their information is derivative, with rather little evidence of critical thinking Most students will fall into this category
D	Minimally acceptable. The students have put in effort but work is marred by some misunderstandings, but not so serious that the student should fail Students falling into this category, and outright failures, would not normally comprise more than about 10%

Source: *Faculty of Arts Handbook*, the University of . . .

What is the problem here? _____

_____

_____

_____

psychology or any other subject just so that students can tell us in their own words what we have told them. We need some sort of 'performance of understanding' (see pp. 74–6) that reflects the kind of understanding that requires an *active demonstration* of the knowledge in question, as opposed to talking or writing about it. This is referred to as 'authentic assessment' (Torrance 1994; Wiggins 1989). The term 'authentic' assessment may imply that all other forms of assessment are inauthentic, so many prefer the term 'performance assessment' (Moss 1992). It reminds us of what we already know in aligned teaching, that the assessment task should require students to do more than

just tell us what they know – unless, of course, declarative knowledge is all that we require in this instance.

Box 9.3 The problem in Box 9.2

The intention is to assess according to quality, but the thinking is still measurement model. Where there is a conflict, it seems that the NRA guidelines would be expected to prevail. For instance, if 30% of students 'showed evidence of original thought as well as having a secure grasp, etc.' that would be seen in this scheme to be anomalous, but as teachers we should be happy if this is what we found. Likewise, we should be disappointed if not ashamed that most students displayed 'derivative information' (C): it looks like they hadn't been taught properly, but here we are told that that is what we should expect. What is wrong here is that the definitions of learning outcome appear to be based on expected distributions of ability. Major departures from that distribution suggest either that there is something wrong with our teaching or that we are too soft in assessing.

Decontextualized assessment

A related issue is whether the assessment tasks should be decontextualized, requiring students to perform in the abstract, out of context. Where the ILOs target declarative knowledge, it is quite appropriate to assess it using decontextualized assessments, such as written examinations. We thus arrive at an important distinction in assessment formats:

1. Decontextualized assessments such as a written exam, or a term paper, which are suitable for assessing declarative knowledge.
2. Performance assessments, such as a practicum, problem solving or diagnosing a case study, which are suitable for assessing functioning knowledge in its appropriate context.

While both decontextualized and contextualized learning and assessment have a place, in practice decontextualized assessment has been greatly overemphasized in proportion to the place declarative knowledge has in the curriculum. As we saw in Chapter 5, functioning knowledge is underwritten by declarative knowledge and we need to assess both. A common mistake is to assess only the lead-in declarative knowledge, not the functioning knowledge that emerges from it. The following ILOs are taken from rehabilitation science, with their SOLO level and type of knowledge assessed:

1. Describe the bones and the muscles of the hand (multistructural, declarative).

Box 9.4 How not to 'mark' a dissertation

A student's postgraduate thesis, carried out at an Australian university, was submitted late, and given a mark of 76. However, during an oral examination, in which the student left the room in tears, one examiner persuaded the other two examiners that because of 'supervisory difficulties', the thesis be upgraded to 79, which meant a classification of second class honours for the degree. The student then raised other issues, including sexual harassment and claimed her thesis was worthy of first class honours. An internal enquiry suggested that 79 be converted to 80, so the dissertation was now awarded first class honours. But the case was then referred to the deputy ombudsman, who advised that the 'real' mark should have been 73, when readjusted for lateness and the bonuses for stress.

A 'real' mark is surely that which reflects the genuine worth of the work done, but here we have a thesis variously marked at 73, 76, 79 and 80, ranging from second to first class honours. The variation is due not so much to differences in staff opinion on the intrinsic academic worth of the thesis, as to differences in opinion on non-academic matters – lateness, stress, supervisory difficulties and sexual harassment – which were factored in arbitrarily and after the event. The public, employers, other universities – not to mention the poor student – would simply have no idea whether the thesis demonstrated those qualities of flair and originality that are associated with first class honours or of the less dazzling but high competence that is associated with good second class honours. It is ironic that a lay person, the deputy ombudsman, seems to have been the one who was least swayed by non-academic issues.

Source: 'From a flood of tears to scandal', *The Australian*, 26 January 2001: p. 4

2 Explain how the bone and muscle systems interact to produce functional movement of the hand, for example in picking up a small coin from the floor (relational, but still declarative).

3 Given a trauma to one muscle group *(x)* rendering it out of action, design a functional prosthesis to allow the hand to be used for picking up a coin (relational, functioning).

Holistic and analytic assessment

Analytic marking of essays or assignments is a common practice. The essay is reduced to independent components, such as content, style, referencing, argument, originality, format, and so on, each of which is rated on a separate scale. The final performance is then assessed as the sum of the separate

ratings. This is very helpful as *formative* assessment (Lejk and Wyvill 2001a); it gives students feedback on how well they are doing on each important aspect of the essay, but the *value* of the essay is how well it makes the case or addresses the question as a whole. The same applies to any task: the final performance, such as treating a patient or making a legal case, makes sense only when seen as a whole.

A valid or authentic assessment must be of the total performance, not just aspects of it. Consider this example from surgery. You want to be sure that the student can carry out a particular operation with high and reliable competence. An analytic assessment would test and mark knowledge of anatomy, anaesthesia, asepsis and the performance skills needed for making clean incisions and then add the marks to see if they reach the requisite 50% (or in this case perhaps 80%). Say a student accrues more than the number of marks needed to pass but removes the wrong part. On the analytic model a pass it must be.

Absurd though this example may seem to be, in an analytic marking scheme some aspects of knowledge are inevitably traded off against others. The solution is not to blur the issue by spreading marks around to fill in the cracks, but to require different levels of understanding or performance, according to the importance of the sub-topic. In this example, the student's knowledge of anatomy was insufficient to allow the correct performance, hence the proper judgment is 'fail'. Assessment of components certainly should be undertaken as formative assessment but, at the end of the road, assessment should address the whole.

In making holistic assessments, the details are not ignored. The question is whether, like the bricks of a building or the characters in a novel, the specifics are tuned to create an overall structure or impact. This requires a *hermeneutic* judgment; that is, understanding the whole in light of the parts. For example, an essay requiring reasoned argument involves making a case, just as a barrister has to make a case that stands or falls on its inherent plausibility. The judge does not judge the barrister's case analytically: uses legal terms correctly (+10 marks), makes eye contacts with jury members (+5 marks), for too long (–3 marks) and then aggregates, the counsel with most marks winning the suit. The argument, as a whole, has to be judged. It is the whole dissertation that passes, the complete argument that persuades, the comprehensive but concise proposal that gets funded, the applicant's case that wins promotion. That is what holistic assessment is about.

Critics argue that holistic assessment involves a 'subjective' judgment. But as we have seen, awarding marks is a matter of judgment too, a series of mini-judgments, each one small enough to be handled without qualm. The numbers make the big decisions: if they add up to 50 or more, then it is a pass. At no point does one have to consider what is the *nature* of a passing grade as opposed to a fail or of a distinction level of performance as opposed to a credit. One of the major dangers of quantitative assessment schemes is that teachers can shelter under them and avoid the responsibility of making the

judgments that really matter: What is a good assessment task? Why is this a good performance? (Moss 1992).

The strategy of reducing a complex issue to isolated segments, rating each independently, and then aggregating to get a final score in order to make decisions, seems peculiar to schools and universities. It is not the way things work in real life. Moss (1994) gives the example of a journal editor judging whether to accept or reject a manuscript on the basis of informed advice from referees. The referees don't give marks, but argue on the intrinsic merits of the paper as a whole and the editor has to incorporate their advice, resolve conflicting advice and make a judgment about what to do with the whole paper: reject it, accept it or send it back for revision. Moss reports that one of her own papers, which argued for a hermeneutic approach to educational assessment, was rejected by the editor of an educational journal on the grounds that a hermeneutic approach was not the model of assessment accepted in the educational fraternity. But it just had been! Moss gleefully pointed out that the editor had used a hermeneutic approach to arrive at that conclusion. Her paper was accepted.

In order to assess learning outcomes holistically, it is necessary to have a conceptual framework that enables us to see the relationship between the parts and the whole. Teachers, like journal editors, need to develop their own framework. The SOLO taxonomy can be useful in assisting that process (see pp. 79–80; Boulton-Lewis 1998; Hattie and Purdie 1998; Lake 1999).

Convergent and divergent assessment: Unintended outcomes

We used the terms 'convergent' and 'divergent' in Chapter 8 in connection with teaching for creativity. A Level 1 view of teaching sees all assessment as convergent: Get right what I have just taught you. When essays are marked with a checklist, marks are awarded only for matching the prescribed points, none for other points that might be just as good or better. This is not what assessment should be about. Virtually all university-level subjects require at least some divergent assessment. Setting only closed questions is like trying to shoot fish in murky water. We need to use open-ended assessment tasks that allow for *unintended outcomes*, that follow from such verbs in the ILOs as 'hypothesize', 'create', 'design', 'reflect' and the like.

A student teacher provided the following metaphor for assessment:

> When I stand in front of a class, I don't see stupid or unteachable learners, but boxes of treasures waiting for us to open.
>
> (An inservice teacher education student, University of Hong Kong)

What 'treasures' students find in their educational experience is something that can surprise, delight and, of course, disappoint too. When we assess using closed questions something like this occurs:

Teacher How many diamonds have you got?
Student I don't have any diamonds.
Teacher Then you fail!
Student But you didn't ask me about my jade.

Students' treasures need not be just in diamonds. If you only ask a limited range of questions, then you may well miss the jade: the treasure that you didn't know existed because you didn't ask. Of course, if the ILOs are expressed only in diamonds that is one thing, but frequently they are not, or ought not to be if they are.

Any rich teaching context is likely to produce learning that is productive and relevant, but unanticipated. The value of many formal activities lies precisely in the surprises they generate, such as field trips, practica or lab sessions, while informal activities bring about unanticipated learning in infinite ways. The student talks to someone, reads a book not on the reading list, watches a television programme, browses the net, does a host of things that sparks a train of thought, a new construction. Such learning probably will not fit the questions being asked in the exam, but they could nevertheless be highly relevant to the course ILOs. Most if not all important discoveries came about as a result of paying attention to unintended outcomes.

Assessment practices should allow for such rich learning experiences, but rarely do. One psychology professor included the following in the final exam paper: 'Based on the first-year syllabus, set and answer your own question on a topic not addressed in this paper.' Another was: 'Psychology. Discuss.' You had to answer these questions extremely well. He also used the instruction: 'Answer *about* five questions.' The conservative or insecure students answered exactly five. The more daring answered three, even two. They were, of course, the deep learners. Other ways of assessing unintended outcomes are reflective journals, critical incidents and the portfolio. We look at these in due course.

Some may see a problem of 'fairness' here. Shouldn't all students be assessed on their performance in the same task? This complaint has weight only in a norm-referenced context, when you are comparing students with each other. Then, yes, you have to standardize so that all have a fair crack at however many As or HDs have been allocated. In portfolio assessment, however, the complaint is irrelevant. If student A can justify task X as addressing the ILOs, and student B task Y, where is the problem?

> To treat everyone the same when people are so obviously different from each other is the very opposite of fairness.
>
> (Elton 2005 on assessing student learning)

If the ILOs specify creativity and originality and the assessment does not allow for them, now that *is* unfair.

Who takes part in assessing?

Three stages are involved in assessing students' performances:

1 *Setting the criteria* for assessing the work.
2 *Selecting the evidence* that would be relevant to submit to judgment against those criteria.
3 *Making a judgment* about the extent to which these criteria have been met.

Traditionally, the teacher is the agent in all three assessment processes. The teacher decides in advance that the evidence for learning comprises correct answers to a set of questions that again in the teacher's opinion addresses and represents the essential core content of the course and the teacher makes the final judgments on meeting the criteria.

Self-assessment (SA) and *peer-assessment* (PA) usually refer to student involvement in stage (3), but students can and often should be involved in stages (1) and (2) as well. Arguments can be made for all or any of these combinations (Boud 1995; Harris and Bell 1986). Students can be involved in discussing with the teacher what the criteria might be, which need not be the same for all students, as happens in a learning contract system (pp. 220–1). Students can also be involved in (2), that is, as the ones responsible for selecting the evidence to be put up against the criteria, as happens with assessment by portfolio. Finally, students can be involved in making the summative judgment (3). This can be as self-assessment or as peer-assessment and either or both can be used as a teaching/learning activity and as an assessment task. Their judgments may also be included in the final grade. All these possibilities are discussed in due course.

Probably the strongest arguments for self- and peer-assessment are that they provide a TLA that engages crucial and otherwise neglected aspects of student learning:

1 First-hand knowledge of the criteria for good learning. Students should be quite clear about what the criteria for good learning are, but when the teacher sets the criteria, selects the evidence and makes the judgment of the student's performance against the criteria, the students may have little idea as to what they should have been doing and where they went wrong. It is too easy for the students just to accept the teacher's judgment and not reflect on their own performance. They should be more actively involved in knowing what the criteria really mean. They should learn how to apply the criteria, to themselves and to others.
2 What is good evidence for meeting the criteria and what is not? Telling students may not engage them. They need to learn what is good evidence being themselves actively involved in selecting it.
3 Making judgments about whether a performance or product meets the given criteria is vital for effective professional action in any field. Professionals need to make these judgments about their own performance (SA) and that of others (PA). It is the learning experience professionals say is

most lacking in their undergraduate education (Boud 1986). Brew (1999) argues that students need to distinguish good from poor information now they are faced with an incredible overload of information from the net: an essential skill in lifelong learning (pp. 148–51). A more general argument along these lines is that conventional assessment disempowers learners, whereas education is about empowering learners and assessment can be made to play an empowering role (Leach et al. 2001).

Reliability and validity

A frequent criticism of qualitative assessment is that it is 'subjective' and 'unreliable'. This is the measurement model talking. Let us rephrase so that it applies to both models of assessment: Can we rely on the assessment results – are they reliable? Are they assessing what they should be assessing – are they valid?

Can we rely on the assessment results?

In the measurement model, reliability means:

- *Stability*: a test needs to come up with the same result on different occasions, independently of who was giving and marking it. Hence, procedure of test–retest reliability: give the same test to the same group again and see if you get the same result.
- *Dimensionality*: the test items need to measure the same characteristic, hence the usual measures of reliability: split-half, internal consistency (Cronbach α).
- *Conditions of testing:* each testing occasion needs to be conducted under standardized conditions.

Here reliability is seen as a property of the test. Such tests are conceived, constructed and used within a sophisticated framework of parametric statistics, which requires that certain assumptions be met, for example that the score distributions need to be normal or bell shaped.

In the standards model reliability means something rather different:

- *Intra-judge reliability.* Does the same person make the same judgment about the same performance on two different occasions?
- *Inter-judge reliability.* Do different judges make the same judgment about the same performance on the same occasion?

Here reliability is not a property of the test, but of the ability of teachers/judges to make consistent judgments. This requires that they know what their framework of judgment is and how to use it: the criteria need spelling out in what are now known as grading criteria or *rubrics*, which are simply clear criteria of grading standards. We deal with these in Chapters 10 and 11.

Reliability here is not a matter of statistical operations, but of being very clear about what we are doing, what learning outcomes we want, what is to be the evidence for those outcomes and why. In other words, reliable assessments are part and parcel of good teaching. We have been explicating the framework and the specific criteria for making informed and reliable judgments about students' learning from Chapter 5 onwards.

Do the test scores assess what they should be assessing?

In the measurement model, the test needs to be validated against some external criterion to show that the trait being measured behaves as it should if it were being measured accurately. Thus, the scores could be correlated with another benchmark test or used as a variable in an experimental intervention, or in predicting an independent outcome.

In the case of the standards model, by way of contrast, validity resides in the *interpretations and uses* to which test scores are put (Messick 1989), that is, in the test's alignment with the total teaching context. For example, if sitting an exam results in students rote-learning model answers, then that is a consequence that invalidates the test. An aligned, or properly criterion-referenced assessment task is valid, a non-aligned one is invalid. The glue that holds the ILOs, the teaching/learning environment, and the assessment tasks and their interpretation together is, again, *judgment*. There is now quite a good deal of agreement about reliability and validity in qualitative assessment (Frederiksen and Collins 1989; Moss 1992, 1994; Shepard 1993; Taylor 1994).

Table 9.2 draws all these points together, contrasting the measurement and standard models.

Task 9.3 (p. 191) is a reflective exercise to help you see where you stand in your thinking about your assessment practice.

Table 9.2 Comparing the measurement and standards models

	Measurement model	*Standards model*
Theory	Quantitative. Classic test theory, using assumptions of parametric statistics	Qualitative. A theory of learning enabling consistent judgments. No assumptions about distributions
Stability	Scores remain stable over testing occasions	Scores after teaching should be higher than before teaching
Dimensionality	The test is unidimensional. All items measure the same construct	Test multidimensional (unless there is only one ILO) The items address all the course ILOs
Testing conditions	Conditions need to be standard	Conditions reflect an individual's optimal learning in the intended application of the learning

(Continued)

Table 9.2 Continued

	Measurement model	*Standards model*
Validity	External: how well the test correlates with outside performances	Internal: how well scores relate to the ILOs and to the target performance domain
Use	Selecting students. Comparing individuals, population norms. Individual diagnosis	Assessing the effectiveness of learning, during and after teaching and learning

Now take a second look at Task 9.1 (p. 165). Would you make different decisions now?

Answers to Task 9.2 The NRA/CRA problem

Despite the fact that Susan's and Robert's performances were compared, the purpose of comparing was not to award the grades but to check the consistency of making the judgment. What happened here was that the initial judgment of Robert's performance was inaccurate, very possibly because of a halo effect: 'Ah, here's Robert's little effort. That won't be an A!' It took a direct comparison with Susan's effort to see the mistake. The standards themselves were unaltered.

Summary and conclusions

Formative and summative assessment

The first thing to get right is the reason for assessing. There are two paramount reasons that we should assess: formative, to provide feedback during learning; and summative, to provide an index of how successfully the student has learned when teaching has been completed. Formative assessment is basic to good teaching, and has been addressed in earlier chapters. Our main concern in this chapter is with summative.

Effects of assessment on learning: Backwash

The effects of assessment on learning are usually deleterious. This is largely because assessment is treated as a necessary evil, the bad news of teaching and learning, to be conducted at the end of all the good stuff. Students second-guess the assessment and make that their syllabus, and will under-

Task 9.3 Where does your assessment stand?

Reflect on your assessment practice so far, put a cross on the continuum on a point that best represents what you currently do in assessing your students:

Formative	_____	Summative
Involving your students	_____	All teacher controlled
Using open-ended assessment tasks	_____	Using closed-ended assessment tasks
Authentic tasks	_____	Decontextualized tasks
Criterion-referenced	_____	Norm-referenced
Using grading criteria	_____	Using model answers
Awarding grades for quality	_____	Awarding marks for quantity
Assessing the task as a whole	_____	Assessing individual components of the task

If you were to adopt constructively aligned assessment, what changes would you need to make in your assessment practice?

estimate requirements if the assessments tasks let them, so they get by with low-level, surface learning strategies. In aligned teaching, contrariwise, the assessment reinforces learning. Assessment is the senior partner in learning and teaching. Get it wrong and the rest collapses. This and following chapters aim to help us get it right.

Measurement model of assessment

The measurement model of educational assessment was hijacked from individual differences psychology, which is concerned with measuring stable characteristics of individuals so that they can be compared with each other or with population norms. However, when this model is applied to assessing educational outcomes, numerous problems arise. Unfortunately, many procedures deriving from the measurement model are incompatible with constructive alignment but remain in current practice: grading on the curve so that students have to compete for the higher grades; marking, despite its universality, has implications for the nature of knowledge that are unacceptable; separating assessment from teaching, which ignores alignment and imposes a separate culture of assessment as apart from the culture of teaching and learning. The backwash from the measurement model sends unfortunate messages to students about the nature of knowledge and about assessment preparation strategies that lead to surface learning.

Standards model of assessment

The standards model of educational assessment defines forms of knowledge to be reached at the end of teaching, expressed as various levels of acceptability in the ILOs and grading system. This framework requires higher levels of judgment on the part of the teacher as to how well the students' performances match the ILOs than does quantitative assessment. The assessment tasks need to be 'authentic' to the ILOs, stipulating a quality of performance that the assessment tasks demand. The backwash tells students they need to match the target performances as well as they are able.

Norm- and criterion-referenced assessment: Let's get it straight

Although norm- and criterion-referenced assessment are logically different, there is still room for confusion, which we try to dispel with some exercises.

Some important concepts in assessment

We present a list of concepts that are important in thinking about and implementing constructive alignment. Authentic assessment directly engages the student with functioning knowledge in its context, decontextualized assessment is more suitable for declarative knowledge. While formative feedback often should be analytic by informing students how well they are managing different aspects of the task, the summative judgment should be of the whole, not the sum of its parts. Open-ended assessment tasks allow for unintended and divergent outcomes, and students themselves need to be involved in the various stages of assessment, in both peer- and self-assessment.

Reliability and validity

Measurement modelists accuse qualitative assessment methods of being 'subjective' and 'unreliable'. What they fail to recognize is that reliability and validity are not the exclusive domains of number crunchers. As the quantitative scaffolding is dismantled, we find that notions as to reliability and validity depended more and more on the teacher's basic professional responsibility, which is to make judgments about the quality of learning.

Further reading

Dart, B. and Boulton-Lewis, G. (eds) (1998) *Teaching and Learning in Higher Education*. Camberwell: Australian Council for Educational Research.

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

Taylor, C. (1994) Assessment for measurement or standards: The peril and promise of large scale assessment reform, *American Educational Research Journal*, 31: 231–62.

Torrance, H. (ed.) (1994) *Evaluating Authentic Assessment: Problems and Possibilities in New Approaches to Assessment*. Buckingham: Open University Press.

The Taylor and Moss articles are seminal, outlining the principles of the rethink on assessment, where the criteria that are qualitatively defined are included. Taylor traces the historical and conceptual roots of NRA and CRA, clearly outlining where the confusions in current practice have crept in, while Moss goes into the conceptual issues in terms of assessment theory. Torrance's book contains some commentaries on the new approach. Dart and Boulton-Lewis contains chapters by Boulton-Lewis, Dart, and Hattie and Purdie, which specifically deal with SOLO as a conceptual structure for holistic assessment.

Websites

University of Melbourne, see especially the Assessment in Australian Universities project: www.cshe.unimelb.edu.au/assessinglearning

The Higher Education Academy: www.heacademy.ac.uk/default.htm

Oxford Brookes University: www.brookes.ac.uk/services/ocsd/2_learntch/2_learnt.html

The Hong Kong Polytechnic University's Assessment project, see especially the Assessment Resource Centre: www.assessment.edc.polyu.edu.hk/

Queensland University of Technology: www.tedi.uq.edu.au/teaching/index.html. Click 'Assessment' and choose your topic.