Teaching for Quality Learning at University
Assessing for learning quality: II. Practice

John Biggs

In this chapter we look at implementing assessment package 2. What assessment tasks are available, and for what purpose is each best used? How can large classes be assessed effectively? How can students be quickly provided with feedback particularly in large classes? When, and how, should self/peer-assessment be used? How can qualitative assessments be combined across several tasks, or across units, to yield a single final grade? How can students' performance be graded qualitatively when the results have to be reported in percentages? These are the bread-and-butter questions we address in this chapter.

What are the best formats for summative assessment?

Let us say you chose assessment package 2 (if you didn't, you might as well skip the rest of this chapter). You are now faced with assessing a large class. I will put it to you in the form of a multiple-choice test item:

My question: What format will you use to assess your class of 400 first-year (biology) students?

1. An individual research project (maximum 5000 words).
2. **A** multiple-choice test.
3. A 2000 word assignment during the term, and a final three-hour examination.
4. A contextualized problem-based portfolio.

Your reply. Not 1, it takes too long to mark; same for 3. In 4 is Biggs trying to be funny, or is he serious but hopelessly unrealistic? Should be 2, which is what most people use, but it's clear what the prejudices of He Who Set the Question are. But I'll risk it and say 2.

Well, you could be right, but the question is unanswerable as it stands. A crucial consideration has been omitted: *what are your objectives?* The 'best' assessment method is the one that best realizes your objectives. In your first year class, are you targeting declarative knowledge, or functioning knowledge, or both? What levels of understanding do you require, and or what topics: knowledge of terminology, description, application to new problems …? As you rightly said in response to our multiple choice question, multiple-choice is widely uses, and yes, it is convenient. But will it assess what you are after?

We need to clarify further. Although you chose package 2, some issues are not entirely clear-cut. Let me again think aloud on your behalf:

- *NRA or CRA?* CRA. I want the grades to reflect learning, not relatives between students. (However, there's no room in second year for all of them, we may have to cull somehow..)
- *Quantitative or qualitative?* Qualitative, I hope, but aren't there certain basic facts and skills I want students to get correct?
- *Holistic or analytic?* Holistic, but how do I combine holistic assessments of several tasks to make one final grade?
- *Convergent or divergent?* Do I want students to get it right, or to show some lateral thinking? Probably both.
- Contextualized or decontextualized? Both. Students must understand the literature, but they need to solve problems in context.
- *Teacher assessed or self/peer assessed?* I intend to be the final arbiter, but self/peer assessment has educational and workload advantages.
- *Backwash?* What effect will my assessment tasks have on students' learning?
- *Time-constrained? Invigilated?* Does my institution require me to impose formal examinations conditions?

There are no right answers, only better or worse ones, and the range of assessment formats to choose from is large. We have to strike a balance between practicality and validity. Chapter 8 set a stern example to live up to, but we have to be realistic. There are 400 students to assess, and their results have to be sent to the board of examiners the week following the examination.

Throughout this chapter, we will be reviewing many different modes of assessment. You should read reflectively as before, with a particular problem class in mind. Ask yourself: how might this help in developing my own assessment practices? At the end of the chapter, we return to the problem posed by the first-year class.

How important is the format of assessment?

First, let us see if it is matters, apart from convenience, whether you use multiple-choice, or essay exam, or assignment. This depends on the activities an assessment format usually elicits. Are they ones the match your teaching objectives? If they do match your objectives, the backwash is positive, but if they do not, the backwash will encourage students to use surface approaches to learning.

The evidence is very clear that different formats do produce typical forms of backwash. They get students doing different things in preparing for them, some being much more aligned to the unit objectives than others. Tang (1991) used questionnaire and interview to determine how physiotherapy students typically prepared for short essay examinations and for assignments (see Box 9.1)

Box 9.1: Learning activities reported by students in preparing for (a) short essay question examination, and (b) assignment

(a) Short essay examination

rote learning, question spotting, going through past papers, underlining, organizing study time and materials, memorizing in meaningful context, relating information, visualizing patients' conditions, discussing with other students.

(b) Assignment

choosing easy questions/interesting questions/what lecturers expect, copying sources, reading widely/searching for information sources, relating question to own knowledge, relating to patients' conditions and clinical application, organizing, revising text to improve relevance, discussing with other students.

Source: from Tang 1991.

In essence, exams tended to elicit memorization-related activities, assignments application-related activities. The assignment required deep learning from the students with respect to one topic, the exam required deep learning from the students with respect to one topic, the exam required acquaintance with a range of topics. The teachers concerned realized that the assignment better addressed the desired course objectives, but only with respect to one topic. They accordingly adopted a policy to use both: short answer exams to ensure coverage, the assignment to ensure depth. A not unusual compromise.

Scouller (1996, 1998) found that students were likely to employ surface strategies in the multiple-choice (MC) format; they saw MC tests as requiring low cognitive level processes. Indeed, Scouller found that using deep approaches was negatively related to MC test performance. The opposite occurred with essays. Students saw essays as requiring higher level processes, and were more likely to use them, and those who didn't, using surface approaches instead, did poorly. Students who preferred MC to essay assignment gave surface-type reasons: you can rely on memory, you can 'play the game' (see Box 9.2). Yet these were the same reasons why other students disliked the MC; these students were angry at being assessed in a way that they felt did not do justice to their learning. When doing assignments, they felt they were able to show higher levels of learning. Short answer examinations did not attract their anger, but the level of cognitive activities assessed was no better than with MC.

Box 9.2: Two examples of students' views on multiple choice tests

I preferred MCQ … It was just a matter of learning facts… and no real analysis or critique was required which I find tedious if I am not wrapped in the topic. I also dislike structuring and writing and would prefer to have the answer to a question there in front of me somewhere.

….A multiple choice exam tends to examine too briefly a topic, or provide overly complex situations which leave a student confused and faced with 'eenie, meenie, minie, mo' situation. It is cheap, and in my opinion ineffectual in assessing a student's academic abilities in the related subject area.

Source: from Scouller 1997.

Assessment by portfolio leads students to see it as 'a powerful learning tool….', and as requiring them to be divergent: 'it led me to think many questions that I never think of' (see p. 136). Wong (1994) used SOLO to structure a secondary 5 (Year 10) mathematics test in the ordered outcome format (see below), and compared students' problem-solving methods on that with those they used on the traditional format. The difference was not on items correct, but on how they went about the problems. They behaved like 'experts' on the SOLO test, solving items from first principles, while on the traditional test they behaved like 'novices', applying the standard algorithms.

In sum then, MCs and short answers tend to elicit low-level verbs, leaving students feeling that MCs and short answers do not reveal what they have learned, while portfolios and SOLO encourage high-level verbs. Unfortunately, there appears to be little further research on backwash from other assessment modes. Tang's study suggests how one might go about this, matching verbs denoted as desirable in the objectives with the verbs students say the assessment tasks encouraged them to use.

We now review particular assessment formats in detail, under four headings: extended prose, objective, performance and rapid assessments, which are particularly suitable for large classes.

Extended prose (essay type) formats of assessment

The essay, as a continuous piece of prose written in response to a question or problem, is commonly intended for assessing higher cognitive levels. There are many variants:

- The timed examination, students having no prior knowledge of the question;
- The open-book examination, students usually having some prior know- and being allowed to bring reference material into the exam room;
- The take-home, where students are given notice of the questions and several days to prepare their answers in their own time;
- The assignment, which is an extended version of the take-home, and comprises the most common of all methods of evaluating by essay;
- The dissertation, which is an extended report of independent research.

Let us discuss these.

Essay examinations

Essay exams are best suited for assessing declarative knowledge. They are usually decontextualized, students writing under time pressure to demonstrate level of their understanding of core content. The format is open-ended, so theoretically students can express their own constructions and views, supporting them with evidence and original arguments. The reality is often different.

The time constraint for writing exams may have several reasons:

1. *Convenience.* A time and a place is nominated for the final assessment, It teachers, students and administration can work around. We all know where we stand.
2. *Invigilation.* Having a specified time and place makes it more easy for the time-keeper to prevent cheating. This enables the institution to guarantee authenticity of the results.
3. *Conditions are standardized.* No one has an 'unfair advantage'. But do you allow question choice in a formal examination? If you do, you violate the standardization condition, because all candidates are not then sitting the 'same' examination (Brown and Knight 1994). Standardization is in fact a hangover from the measurement model; it is irrelevant in a criterion-referenced situation.
4. *Models real lift* The time constraint reflects 'the need in life to work swiftly, under pressure and well' (Brown and Knight 1994: 69). This is unconvincing. In real-life situations where functioning knowledge is time-stressed — the operating theatre, the bar (in the courts, that is) or classroom — this point is better accommodated by performance assessment, rather than by pressurizing the assessment of declarative knowledge in the exam room. Alignment suggests that time constraints be applied only when the target performance is itself time-constrained.

Time constraint creates its own backwash. Positively, it creates a target for students to work towards. They are forced to review what they have leamed throughout the unit, and possibly for the first time see it as a whole -tendency greatly enhanced if they think the exam will require them to demonstrate their holistic view. Students' views of examinations suggest that this rarely happens.

The more likely backwash is negative; students memorize specific points to be recalled at speed (Tang 1991). Students go about memorization differently. Learners who prefer a deep approach to learning create a structure first, then memorize the key access words ('deep-memorizing') while surface learners simply memorize unconnected facts (Tang 1991). So while timed exams encourage *memorizing,* this is not necessarily rote memorizing or surface learning. Whether it is or not depends on students' typical approaches to learning, and on what they expect the exam questions to require.

Does the time constraint impede divergent responses? Originality it is a temperamental horse, unlikely to gallop under the stopwatch. However, if students can guess likely questions, they can prepare their original at leisure; and with a little massaging of the exam question, express prepared creation. You as teacher can encourage this high-level off-track preparation, by making it known you intend asking very open questions ('What is the most important topic discussed in the unit this semester? Why?", or by telling the students at the beginning of the semester what the exam questions will be. Assessing divergent responses must be done holistically. The use of a model answer checklist does not allow for the well argued surprise. Students should be told how the papers are to be marked; then they can calculate their own risks.

In sum, time constraints in the exam room cannot easily be justified educationally. The most probable effect is to encourage memorization, with or without higher-level processing. In fact, time constraints exist for administrative not educational reasons. They are convenient, and they make cheating more difficult. Whether these gains are worth the educational costs is a good question.

Open-book examinations remove the premium on memorization of detail, but retain the time constraint. Theoretically, students should be able to think about higher-level things than getting the facts down. Practically, they need to be very well organized; otherwise they waste time tracking down too many sources.

Exams are almost always teacher assessed, but need not be. The questions can be set in consultation with students, while the assessing and award of grades can be done by the students themselves, and/or their-peers, as we saw in Chapter 8. The backwash, and range of activities being assessed, change dramatically with self/peer assessment.

The assignment, the term-paper, the take-home

The assignment or term paper, deals with declarative knowledge, the project (see below) with 'hands-on' research-type activities. The assignment is not distorted by immediate time limitations, or by the need to rely on memory. In principle, it allows for deeper learning; the student can consult more sources and, with that deeper knowledge base, synthesize more effectively. However, plagiarism is easier, which is why some universities require that a proportion of the assessments in a unit are invigilated. The take-home with shorter time limits, often overnight, makes plagiarism a little more difficult.

Self/peer-assessment can be used to assess assignments. Given the criteria, the students award a grade (to themselves, to a peer's paper or both), and justify the grade awarded. That in itself is a useful learning experience. But whether the self/peer grading(s) stand as the official result, or part of it, are matters that can be negotiated. In my experience, students like the self-assessing process, but tend to be coy about its being a significant part of the result.

Assessing extended prose

Years ago, Starch and Elliot (1912; Starch 1913a,b) originated a devastating series of investigations into the reliability of assessing essays. Marks for the same essay ranged from bare pass to nearly full marks. Sixty years later, Diederich (1974) found things just as bad. Out of the 300 papers he received in one project, 101 received every grade from 1 to 9 on his nine-point marking scale.

Judges were using different criteria. Diederich isolated four families of criteria, with much disagreement as to their relative importance:

Ideas: originally, relevance, logic.
Skills: the mechanics of writing, spelling punctuation, grammar.
Organization: format, presentation, literature review.
Personal style: flair.

Each contains a family of items, according to subject. 'Skills' to Diederich meant writing skills, but they could be 'skills' in mathematics, chemistry or fine arts. Likewise for the other components: ideas, organization and personal style. It would be very valuable if staff in a department collectively clarified what they really are looking for under these, or other, headings.

Back to the holistic/analytic question

When reading an essay, do your rate separately for particular qualities, such as those mentioned by Diederich, and then combine the ratings in some kind of weighted fashion? Or do you read and rate the essay as a whole, and give an overall rating?

We dealt with the general argument in Chapter 8. The analytic method of rating the essay on components, and adding the marks up, is appealing. It leads to better agreement between markers. But it is slow. Worse, it does not address the essay as a whole. The unique benefit of the essay is to see if students can construct their response to a question or issue framework set by the question. They create a 'discourse structure', which is the point of the essay. Analytic marking is ill-attuned to appraise discourse structure.

Assessing discourse structure requires a framework within which that holistic judgement can be made. SOLO helps you to judge if the required structure is present or not. Listing, describing and narrating are structural structures. Compare-and-contrast, causal explanation, interpretation and so on are relational. Inventive students create their own structures, which when they work can make original contributions; are extended abstract.

The facts and details play their role in these structures in like manner to the characters in a play. And the play's the thing. You do not ignore details, but ask of them:

- Do they make a coherent structure (not necessarily the one you had in mind)? If yes, the essay is at least relational.
- Is the structure the writer uses appropriate or not? If yes, then the question has been properly addressed (relational). If no, you will have to decide how far short of satisfactory it is.
- Does the writer's structure open out new ways of looking at the issue? If yes, the essay is extended abstract.

If the answer is consistently 'no' to all of the above, the essay is multi-structural or less, and should not be given good marks, because that is not the point of the essay proper. If you do want students to list points, the short answer, or even the MC, is the appropriate format. These are easier for the student to complete, and for you to assess.

This distinction recalls that between 'knowledge-telling' and 'reflective writing' (Bereiter and Scardamalia 1987). Knowledge-telling is a multi-structural strategy that can all too easily mislead assessors. Students focus only on the topic content, and tell all they know about it, often in a listing or point-by-point form. Using an analytic marking scheme, it is very hard not to award high marks, when in fact the student hasn't even addressed the question. Take this example of an ancient history compare-and-contrast question: 'In what ways were the reigns of Tutankhamen at Akhnaton alike, and in what ways were they different?' The highest scoring student gave the life histories of both pharaohs, and was commended on her effort and depth of research, yet her discourse structure was entirely inappropriate (Biggs 1987b).

Reflective writing transforms the writer's thinking. E. M. Forster put it thus: 'How can I know what I think until I see what I say?' The act of writing externalizes thought, making it a learning process. By reflecting on what **you see,** you can revise it in so many ways, creating something quite new, even to yourself. That is what the best academic writing should be doing.

The essay is obviously the medium for reflective writing, not knowledge-telling. Tynjala (1998) suggests that writing tasks should require students;

- actively to transform their knowledge, not simply to repeat it;
- to undertake open-ended activities that make use of existing knowledge-beliefs, but that lead to questioning and reflecting on that knowledge;
- to theorize about their experiences;
- to apply theory to practical situations, and/or to solve practical problems or problems of understanding.

Put otherwise, the question should seek to elicit higher relational and extended abstract verbs. Tynjala gave students such writing tasks, which they discussed in groups. They were later found to have the same level of edge as a control group, but greatly exceeded the latter in the *use* to which they could put their thinking. The difference was in their functioning in their declarative, knowledge.

Maximizing stable essay assessment

The horrendous results reported by Starch and Elliott and by Diederich occurred because the criteria were unclear, were applied differently by different assessors and were often unrecognized. The criteria must be aligned to the objectives from the outset, and be consciously applied.

Halo effects are a common source of unreliability. Regrettable it may be, but we tend to judge the performance of students we like more favourably than those we don't like. Attractive female students receive significantly higher grades than unattractive ones (Hore 1971). Halo effects also occur order in which essays are assessed. The first half-dozen scripts tend to set standard for the next half-dozen, which in turn reset the standard next. A moderately good essay following a run of poor ones tends to be assessed higher than it deserves, but if it follows a run of very good ones, it is marked down (Hales and Tokar 1975).

Halo and other distortions can be greatly minimized by discussion; judgements are social constructions (Moss 1994; see pp. 81, 99 above). There is some really strange thinking on this. A common belief is that it is more objective' if judges rate students' work without discussing it. In one fine arts department, a panel of judges independently award grades without discussion; the student's final grade is the undiscussed average. The rationale for this bizarre procedure is that works of an artist cannot be judged against outside standards. Where this leaves any examining process I was unable to discover.

Out of the dozens of universities where I have acted as an external examiner for research dissertations, only one invites examiners to resolve disagreement by discussion before the higher degrees committee adjudiates. Consensus is usually the result. Disagreements between examiners are more commonly resolved quantitatively: for example, by counting heads, or by hauling in additional examiners until the required majority is obtained. In another university I could mention, such conflicts are resolved by a vote in senate. The fact that the great majority of senate members haven't seen the thesis aids detachment. Their objectivity remains unclouded by mere knowledge.

Given all the above, the following precautions suggest themselves:

- All assessment should be 'blind', with the identity of the student concealed.
- All rechecking should likewise be blind, with the original mark concealed.
- Each question should be marked across students, so that a standard for each *question* is set. Marking by the student rather than by the question allows more room for halo

effects, a high or low mark on one question influencing your judgement on the student's answers to other questions.

- Between questions, the papers should be shuffled to prevent systematic order effects.
- Grade coarsely (qualitatively) at first, say into 'excellent', 'pass' and 'fail', or directly into the grading categories. It is then much easier to discriminate more finely within these categories.
- Departments should discuss standards, to seek agreement on what constitutes excellent performances, pass performances and so on, with respect to commonly used assessment tasks.
- Spot-check, particularly borderline cases, using an independent assessors. Agree on criteria first.
- The wording of the questions should be checked for ambiguities by a colleague.

Objective formats of assessment

The objective test is a closed or convergent format requiring one correct answer. It is said, misleadingly, to relieve the marker of 'subjectivity" in judgement. But judgement is ubiquitous. In this case, it is simply shifted from scoring items to choosing items, and to designating which alternatives are correct. Objective testing is not more 'scientific', or less prone to error. The potential for error is pushed to the front end, where the hard work is designing and constructing a good test. The advantage is that the cost-benifits rapidly increase the more students you test at a time. With machine scoring, it is as easy to test one thousand and twenty students as it is to test twenty: a seductive option.

The following forms of the objective test are in common use:

- Two alternatives are provided (true—false).
- Several, usually four or five, alternatives are provided (the MC).
- Items are placed in two lists, and an item from list A has to be matched an item from list B (matching).
- Various, such as filling in blank diagrams, completing sentences. One version, the cloze test, is used as a test of comprehension.
- Sub-items are 'stepped' according to difficulty or structure, the student being required to respond as 'high' as possible (the ordered outcome).

Of these, we now consider the MC, and the ordered outcome. The cloze is considered later, under 'rapid' assessment.

Multiple-choice tests

The MC is the most widely used objective test. Theoretically, MCs can assess high-level verbs. Practically, they rarely do, and some students, the Susans rather than the Roberts, look back in anger at the MC for not doing so (Scouller 1997). MCs assess declarative knowledge, usually in terms of the least demanding process, recognition. But probably the worst feature of MCs is that they encourage the use of game-playing strategies, by both student and teacher. Some examples:

Student strategies

- In a four-alternative MC format, never choose the facetious or the jargon-ridden alternatives.
- By elimination, you can create a binary choice, with the pig-ignorant having a 50 per cent chance of being correct.
- Does one alternative stimulate a faint glow of recognition in an otherwise unrelieved darkness? Go for it.
- Longer alternatives are not a bad bet.

Teacher strategies

- Student strategies are discouraged by a guessing penalty: that is, deducting wrong responses from the total score. (Question: why should this be counter-productive?)
- The use of facetious alternatives is patronizing if not offensive (I can play games with you but you can't with me). Not nice.
- Rewording existing items when you run out of ideas. Anyway, it increases reliability.

MC tests have great coverage, that 'enemy of understanding' (Gardner 1993). One hundred items can cover an enormous number of topics. But if there is exclusive use of the MC, it greatly misleads as to the nature of knowledge, because the method of scoring makes the idea contained in any one item the same value as that in any other item. But consider Lohman's (1993) instance, where an MC test was given to fifth-grade children on the two hundredth anniversary of the signing of the US Constitution. The only item on the test referring to Thomas Jefferson was: 'Who was the signer of the Constitution who had six children?' A year later, Lohman asked a child in this class what she remembered of Thomas Jefferson. Of course, she remembered that he was the one with six children, nothing of his role in Constitution. Students, including tertiary students, quickly learn that 'There is no need to separate main ideas from details; all are worth one point. And there is no need to assemble these ideas into a coherent summary or to integrate them with anything else because that is not required' (Lohman 1993: 19). The message is clear. Get a nodding acquaintance with as many details as you can, but do not be so foolish as to attempt to learn anything in depth.

MC tests can be useful if they supplement other forms of assessment, but when used exclusively, they send all the wrong signals. Unfortunately, they *are* convenient.

Ordered outcome items

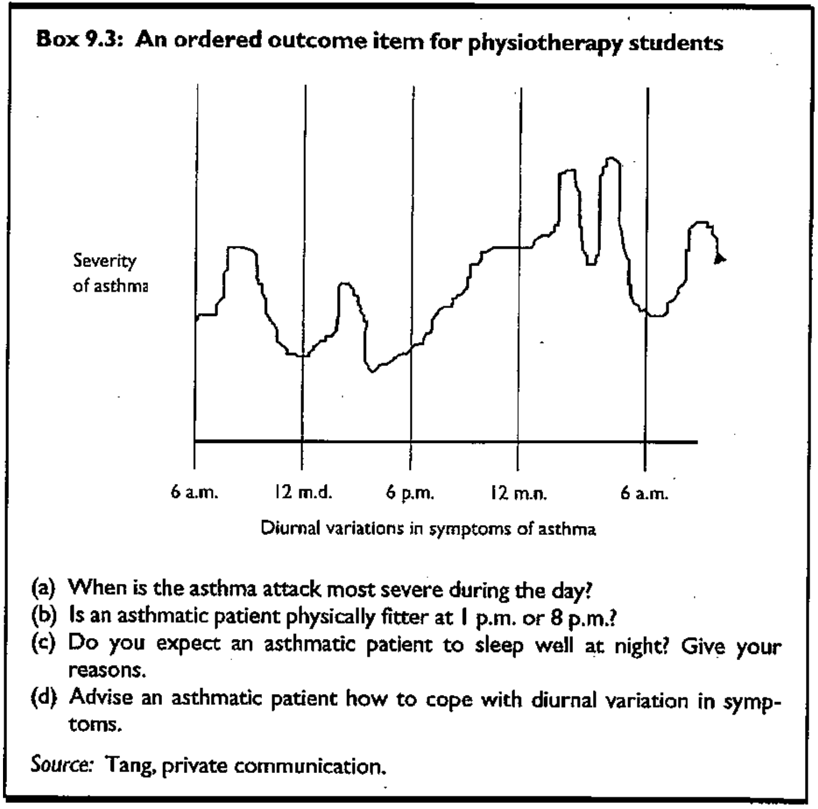
An ordered outcome item looks like an MC, but instead of opting for the one correct alternative out of the four or so provided, the student is required to attempt all sub-items (Masters 1987). The sub-items are ordered into a hierarchy of complexity that reflects successive stages of learning that concept or skill. The students ascend the sequence as far as they can, thus indicating their level of competence in that topic.

All that is required is that the stem provides sufficient information for a range of questions of increasing complexity to be asked. How those questions are derived depends on your working theory of learning. SOLO can be used as a guide for working a sequence out. A SOLO sequence would look like this:

1. *Unistructural:* use one obvious piece of information coming directly from the stem.
2. *Multistructural:* use two or more discrete and separate pieces of information contained in the stem.

3 *Relational:* use two or more pieces of information each directly related to an integrated understanding of the information in the stem.

4. Extended abstract use an abstract general principle or hypothesis which can be derived from, or suggested by, the information in the stem.



Box 9.3: An ordered outcome item for physiotherapy students

Diurnal variations in symptoms of asthma

(a) When is the asthma attack most severe during the day?
(b) Is an asthmatic patient physically fitter at 1 p.m. or 8 p.m.?
(c) Do you expect an asthmatic patient to sleep well at night? Give your reasons.
(d) Advise an asthmatic patient how to cope with diurnal variation in symptoms.

Source: Tang, private communication.

The student's score is the highest correct level. If the response to the first question is inadequate, the student's understanding is assumed to be prestructural.

The levels do not, however, need to correspond to each SOLO level, or to SOLO levels at all. In a physiotherapy course (C. Tang, private communication), an extended abstract option was inappropriate for the first year, and so two levels of relational were used, as in (c) and (d) in Box 9.3, where (c) refers to conceptual integration (declarative) and (d) to application (functioning). Sub-item (a) is unistructural because it only requires a correct reading of the diagram: a simple but essential first skill. Sub-item a multistructural response, requiring the comparison of two different readings. Sub-item (c) requires interpretation at a simple relational level response, while (d) is relational but more complex, requiring a complete interpretation integrated with functioning knowledge of caring skills.

Key situations can be displayed in this format, and a (d) or (c) level of performance required (in this case, anything less would not be of much help to patients). It is sometimes possible to use a one-correct-answer format for extended abstract items: 'Formulate the general case of the preceding (relational) item is an instance.' Often, however, extended abstract items use open-ended verbs, so we have in effect a divergent short-answer sub-item: 'Give an example where (c) - the preceding item - does *not* occur. Why doesn't it?'

The ordered outcome format sends a strong message to students that higher is better: recognition and simple algorithms won't do. This was the format in which Wong (1994) found students to behave theoretical, like experts do (see p. 168).

Constructing ordered outcome items is the difficult part. The items need to form a staircase: unistructural items must be easier than multtstructal, and multistructural than relational, and relational than extended abstract. This can be tested with trial runs, preferably using the Guttman (1941) scalogram model, or software is available (Masters 1988). Hattie and Purdie (1998) discuss a range of measurement issues involved in the construction and interpretation of ordered outcome SOLO items. Basically, it is as always a matter of judgement.

Scoring ordered outcome items makes most sense on a profile basis. That is, you have nominated key situations or concepts, about which the students need to achieve a minimal level of understanding. In the physio item, (c) is possibly adequate in first year, but by the second year students really should be responding at an applied treatment (d) level. The profile sets minimum standards for each skill or component.

It is tempting to say (a) gets 1 mark, (b) 2 marks, (c) 3 marks, and (d) (let's be generous) 5 marks. We then throw the marks into the pot with all the other test results. However, this destroys the very thing we are trying to assess, a level of understanding. If the score is less than perfect, a nominal understanding of one topic could be averaged with a performative understanding of another, yielding 'moderate' understanding across all topics, which wasn't the case at all.

Performance assessment

Performance assessment requires students to perform tasks that mirror the objectives of the unit. Students should be required to demonstrate that they *see and do things differently* as a result of their understanding.

The problems or tasks set are, as in real life, often divergent or ill-formed, in the sense that there are no single correct answers. For example, there are many acceptable ways a software program should be written for use in an estate agency office. What is important is that the student shows how the problem may reasonably be approached, how resources and data are used, how previously taught material is used, how effectively the solutions meets likely contingencies and so on. Clearly, this needs an open-ended assessment format and assessment process. Almost any scenario from the professions can be used: designing a structure, teaching a new topic, dealing with a patient with a strange combination of symptoms.

Various formats reflect this authentic intention with varying fidelity.

The practicum

The practicum, if properly designed, should call out all the important verbs needed to demonstrate competence in a real-life situation, such as practice teaching, interviewing a patient, any clinical session, handling an experiment in the laboratory, producing an artistic product. It goes without saying that CRA is the most appropriate way of evaluation. An assessment checklist should *not* look like this:

A: Definitely superior, among the best in the year
B: Above average
C: Average
D: Below average, but meets minimal standards
E: Not up to standard.

It shou1d be quite clear that the student has to perform certain behaviours to a specified standard. It then remains to find if the learner can perform them, and if not, why not. Video-taping is useful, as then students can rate their own performance against the checklist of desired behaviours before discussing the supervisor's rating.

The closer the practicum is to the real thing, the greater its validity. The one feature that distorts reality is that it *is* an assessment situation, so that some students are likely to behave differently from the way they would if they were not being observed and assessed. This may be minimized by making observation of performance a continuing fact of life. With plenty of formative assessment before the final summative assessment, the student might nominate when he or she is 'ready' for final assessment. This might seem labour intensive, but recording devices can stand in for *in vivo* observation, as can other students.

In fact, the situation is ideal for peer assessment. Students become accustomed to being observed by each other, and they can receive peer feedback. Whether student evaluations are then used, in whole or in part, in the summative assessment is worth considering. In surgery possibly not; in the expressive arts possibly so.

Presentations and interviews

The class presentation is evaluated in terms of what content is conveyed, and how well. Where the focus is on declarative understanding, the students declaring to their peers, we have the traditional *seminar,* which is not necessarily meant to reproduce a situation in which students will later find themselves. The seminar, if used carefully, offers good opportunities for formative discussion, and peer assessment both formative and summative. However, as we have seen (pp. 86-7 above), it can easily become a poor substitute for proper teaching.

Student presentations are best for functioning rather than declarative knowledge. Peer input can be highly appropriate in this case. The Fine Arts Department at the University of Newcastle (NSW) (not the C mentioned earlier) has an examining panel comprising teachers, a prominent local artist and a student (rotating), who view all the student productions, have a plenary discussion with all staff and students about each, and then submit a final, public, examiners' report. This is not only a very close approximation to real life in the gallery world, but actively involves staff and students in a way that it is rich with learning opportunities.

The *poster presentation* follows the well known conference format. A student or group of students display their work according to an arranged art format during a poster session. This provides excellent opportunities for peer-assessment, and for fast feedback of results. However, Brown and Knight (1994: 78) warn that the poster' must be meticulously prepared'. The specifications need to be very clear, down to the size of the display, and how to use back-up materials: diagrams, flow-charts, photographs. Text needs to be clear and highly condensed. Assessment criteria can be placed on an assessment sheet, which all students receive to rate all other posters. Criteria would include substance, originality, impact and so on.

The *interview* is used most commonly in the examination of dissertations and theses. In the last case, the student constructs a 'thesis' that has to be 'defended' against expert criticism. Almost always, these oral defences are evaluated qualitatively. The student makes a case, and is successful, conditionally successful, unsuccessful but is given another try (with or without formal re-examination) or irredeemably unsuccessful. Here again the criteria are usually clear spelt out: the structure of the dissertation, what constitutes good procedure, what is acceptable and what unacceptable evidence, clarity of writing, format and so on. These criteria are seen as 'hurdles' - they have to be got right eventually -while the assessment itself. is on the *substance* and *originality* of the thesis itself.

In undergraduate teaching, the interview is seen as 'subjective' (which it is, but see above), and it 'takes too long'. However, a properly constructed interview schedule could see a fruitful interview through in 20 minutes, possible 30. How long does it take to assess properly each written product of a three-hour examination, or a 2500 word assignment? Thirty minutes? Gobbets (see below) could be a useful way of structuring and focusing an assessment interview. Unstructured interviews can be unreliable, but bear in mind that the point of interviewing is lost if the interview is too tightly structured.

That point is that the interview is interactive. Teachers have a chance to follow up and probe, and students have a chance to display their jade: their unanticipated but valuable learning treasures. Certainly, the interview might be supplemented with a written MC or short answer (to cover the basics), but the most interesting learning could be brought to light and assessed within 20 minutes or so. Oral assessments should be tape recorded, both in case of dispute (when student and an adjudicator can hear replay) and so that you may assess under less pressure, or subsequently check your **original assessment.**

Self-assessment is an interesting option here, with the teacher- and self-assessments themselves being the subject of the interview.

Critical incidents

Students can be asked to report on 'critical incidents' that seem to them powerful examples of unit content, or that stimulate them to think deeply about the content. They then explain why these incidents are critical, how they arose and what might be done about it. This gives rich information about how students (a) have interpreted what they have been taught, and make use of the information.

Such incidents might be a focus in a reflective journal, or be used as portfolio items (see below).

Project

Whereas an assignment usually focuses on declarative knowledge, the project focuses on functioning knowledge applied to a hands-on piece of research. Projects can vary from simple to sophisticated, and often in the latter case will be best carried out by a group of students. The teacher can allot their respective tasks, or they can work them out among themselves.

There are several ways of awarding grades for a group project. The simplest is to give an overall grade for the project, which each student receives. The difficulty is that it does not allow for passengers, and some of the harder workers will feel aggrieved. Various forms of peer-assessment may be used to modify this procedure, most of which rely on quantification:

- The project is awarded 60 per cent; there are four participants, so there are 240 marks to be allocated. You find out as best you can who did what, and you grade the sections accordingly.

- The project is awarded 60 per cent; there are four participants, so there are 240 marks to be allocated. The students decide who is to get how many marks, with criteria and evidence of effort. One problem is that they may be uncontroversial and divide them equally – some hating themselves as they do so.

- The project is awarded 60 per cent; there are four participants. Each receives a basic 40 per cent. There are now 20 x 4 marks to be allocated. Again, they decide the allocation. The most blatant passenger gets no more, and ends up with 40 per cent; the best contributor gets half of the remainder, by agreement, and ends up with 80 per cent, and so on. This mitigates, slightly, egalitarian pressures.

Some qualitative alternatives:

- Where there is a category system of grading, all receive the same grade.
- The students grade each other, building extent of contribution into the grading system.
- The students grade each other according to contribution, but you decide the categories to be allocated.

A problem with collaborative projects is that individual students too easily focus only on their own specific task, not really understanding other components, or how they contribute to the project as a whole. The idea of a group project is that a complex and worthwhile task can be made manageable, each student taking a section he or she can handle. However, the tasks are divided all too readily according to what students are already good at: Mario will prepare the literature review, Sheila will do the stats. In that case, little *learning* may take place. We want students to learn things other than what they already know, so the allocation might be better be decided so that Sheila does the literature review, and Mario the stats. This is likely to end up with each helping the other, and everyone learns a lot more.

Most importantly, we want them to know what the whole project is about, and how each contribution fits in, so an additional holistic assessment is necessary. For that a concept map would be suitable, or a short answer. And perhaps that is the answer to the group-sharing problem. If a student fails the holistic part, that student fails the project. The backwash is this: make sure you know what your colleagues are doing and why.

Contracts

Contracts replicate a common everyday situation. A contract would take into account where an individual is at the beginning of the course, what relevant attainments are possessed already, what work or other experience, and then, within the context of the course objectives, he or she is to produce a needs analysis from which a programme is negotiated: what is to be done and how it is proposed to do it, and within what time-scale. Individuals, or homogeneous groups of students, would have a tutor to consult throughout, with whom they would have to agree that the contract is met in due course. The assessment problem hasn't gone away, but advantage is that the assessments are tied down very firmly from the add the students know where they stand (Stephenson and Laycock 1993).

A more conventional and less complicated contract is little different from clear criterion-referencing: 'This is what an A requires. If you can prove to me that you can demonstrate those qualities in your learning, then an A is what you will get.' This is basically what is involved in portfolio assessment (see below).

Reflective journal

In professional programmes, it is useful if students keep a reflective journal, in which they record any incidents, thoughts or reflections that are relevant to the unit. Journals are valuable in capturing the students' judgement as to relevance, and their ability to reflect upon experience the content taught. Such reflection is basic to proper professional functioning. The reflective journal, then, is especially useful for assessing content knowledge, reflection, professional judgement and application.

Assessment can be delicate, as journals are often very personal; and boring, as they are often very lengthy. It is a good idea to ask students to submit selections, possibly focusing on critical incidents. Journals should not be 'marked', but taken as evidence of quality in thinking.

Case study

In some disciplines, a case study is an ideal way of seeing how students can apply their knowledge and professional skills. It could be written up as a project, or as an item for a portfolio. Case studies might need to be highly formal and carried out under supervision, or be carried out independently by the student. Possibilities are endless.

Assessing the case study is essentially holistic, but aspects can be used both for formative feedback and for summative assessment. For example, there are essential skills in some cases that must be got right: otherwise the patient dies, the bridge collapses or other mayhem ensues. The component skills here could be pass—fail; fail one, fail the lot (with latitude according to the skill and case study in question). Having passed the components, however, the student then has to handle the case itself appropriately, and that should be assessed holistically.

There are some excellent software options for clinical decision-making for medical case studies, which fit the authentic format extremely well. However, this is a rapidly expanding area and no doubt other disciplines will have their own versions in due course.

Portfolio assessment

In a portfolio, the student presents and explains his or her best 'learning treasures' (p. 155) *vis-à-vis* the objectives. Students have to reflect and use judgement in assessing their own work, and explain its match with the unit objectives. When students give their creativity free reign, portfolios are full of complex and divergent surprises, aligned to the unit aims in ways that are simply not anticipated by the teacher.

In their explanations for their selection of items, students explain how the evidence they have in their portfolios addresses their own or the official unit aims. One danger with portfolios is that students may go overboard, creating excessive workload both for themselves and for the teacher. Limits must be set (see below).

Assessing portfolio items can be deeply interesting. It may be time consuming, but that depends on the nature and number of items. Many items, such as concept maps, can be assessed in a minute or so, In any event, a morning spent assessing portfolios feels like 30 minutes at look-alike assignments. Following are some suggestions for implementing portfolio assessment.

1 *Make it quite clear in the teaching objectives what the evidence for good learning may be.* The objectives should be available to students at the beginning of the semester.
2 *State the requirements for the portfolio.* These need to be made very clear.

- *Number of items.* In a semester-long unit, four items is about the limit.
- *Approximate size of each item.* The total portfolio should not be longer than a project or assignment you would normally set. I suggest no more than 1500 words for any one item, but that depends on the nature of the item. Some items, such as concept maps or other require less than a page.
- *A list of sample items;* but emphasize that students should show some creativity by going outside that list, as long as the items are relevant. Items should not be repetitive, making the same point in different ways.
- *Any compulsory items?* In my courses (in teacher education) I usually prescribe a journal, leaving the other items to student choice.
- *Source of items.* Items may be specific to a unit, or drawn from other units in the case of evaluating at the end of a course/programme. In some problem-based courses, students will be continually providing often on a pass/fail basis, over a year, or two years. The final evaluation could then comprise - *in toto* or in part - samples of the best work tents think they have done to date.
- *What are the items supposed to be getting* at? Are your teaching objectives best addressed as a package, or as a list of separate items?

3 *Decide how the portfolio is to be graded.* There are two alternatives:
  a) assessing individual items, and then combining;
  b) assessing the portfolio as a whole (the ‘package’).

If (a), the situation is the same as combining several assessments within a unit to arrive at a final grade (see pp. 190-4 below). It is tempting to mark each item separately, and then total, but that misses the point of the portfolio, which is embedded in (b). Each item should address some aspect learning, so that the whole addresses the thrust of the unit. This really get back to *your* conception of your unit: do you see yourself teaching a collection of topics, or do those topics constitute a *thrust?* If the latter, the students’ portfolios should address that thrust. In the last case, the student is in effect saying: *'This* is what I got out of your class. I have learned these things, and as a result my *thinking* is changed.' If their package can show that they have learned well indeed.

You might include other assessment tasks apart from the portfolio: for example, a conventional assessment to establish 'coverage' of basics. You will then need to decide how to combine the two sets of results.

Portfolios have been used for years in the fine arts, but they can be used to assess almost any course content. A case study of portfolio assessment is in Chapter 10.

Assessing in large classes

If lecturing is the default for large-class teaching, MCs and timed exams are the default for large-class assessment. Exams take a lot of time to assess, but with tutor assistance and the clear time slots in which things have to be done and reported, we can come to terms with them. Unfortunately, as we have seen, exams are not the best modes of assessment. We now look at alternatives for large-class assessment that:

a) are rapidly administered, completed and assessed;
b) get at higher order learnings than is usually the case with the two default modes.

First some strategic decisions need to be made.

1. You maybe able to justify postponing time-consuming qualitative assessments in the first year, such as individual practica or portfolios, to the, it is hoped, sparer second and third years. At least students will have the experience of these assessments before they do graduate.
2. Cut down on massive, mind-numbing single-mode assessment such as the final exam. Assess more often, with more varying assessments (Brown and Knight 1994; Davis and McLeod 1996b).

Let us then see what further assessment tasks we might use.

Concept maps

Concept maps, introduced as a TLA (see pp. 82-3), can also be used for assessment. They enable us to tell at a glance if a student has an impoverished knowledge structure relating to the topic, or a rich one- (see Task 5.1). While they are to be assessed holistically, you could rate the structure on a 10-point scale, say, just in order to derive a figure for reporting.

Venn diagrams

Venn diagrams are a simple form of concept map, where the boundary of a concept is expressed in a circle or ellipse, and interrelations between concepts are expressed by the intersection or overlap of the circles. Venn diagrams, like concept maps, are very economical ways of expressing relationships. They can be used for teaching purposes, in conveying relationships to learners, and for assessment purposes, so that learners may convey their ways of seeing relationships between concepts. Getting students to draw and briefly explain their own Venns, or to interpret those presented, can be done quickly, where the target of understanding is relationships between ideas. Venns make good gobbets (see below).

Three-minute essay

We met the three-minute essay in Chapter 6, as a method of introducing reflective activity into large-class teaching, by asking such questions as:

- What do I most want to find out in the next class?
- What is the main point I learned today?

These questions may provide very useful information for the teacher in two respects: formatively, in finding out how the content is being interpreted by students; and summatively, in finding out if students have made appropriate interpretations, which can be used for grading purposes. Such questions can be answered in minutes in a large class.

Short answer examinations

In short answers, the student answers in note form. This format is useful for getting at factual material, e.g. addressing or interpreting diagrams, charts and tables, but is limited in addressing main ideas and themes. The examination is usually after something quite specific, and operates in practice more like the objective format than the essay (Biggs 1973; Scouller 1996). However, it has advantages over the standard MC, in that it is less susceptible to test-taking strategies (you can’t work out the answer by elimination), it requires active recall rather than just recognition and it is easier for you to construct, but not as easy to score.

Gobbets

Gobbets are significant chunks of content with which the student should be familiar and to which the student has to respond (Brown and Knight 1994). They could be a paragraph from a novel or of a standard text, a brief passage of music, a Venn diagram, an archaeological artifact, a photograph (a building, an engine part) and so on. The students’ task is to identify the gobbet, explain its context, say why it is important, what it reminds them of, or whatever else you would like them to comment on.

Gobbets should access a bigger picture, unlike short answers, which are it sufficient unto themselves. That big picture is the target, not the gobbet itself. Brown and Knight point out that three gobbets can be completed in the time it takes to do one essay exam question, so that to an extent you can assess both coverage and depth.

Letter to a friend

In the 'letter to a friend', the student tells an imaginary or real friend, who is thinking of enrolling in the unit next year, about his or her own experience of the unit (Trigwell and Prosser 1990). These letters are about a page in length and are written and assessed in a few minutes. The students should reflect on the unit and report on it as it affects them. Letters tend to be either multistructural or relational, occasionally extended abstract. Multistructural letters are simply lists of unit content, a rehash of the course outline. Good responses provide integrated accounts of how the topics fit together and form a useful whole, while the best describe a change in personal perspective as a result of doing the unit. They also provide a useful source of feedback to the teacher on aspects of the unit.

Like the concept map, letters supplement more fine-grained tasks with an overview of the unit, and they make good portfolio items

Cloze tests

Cloze tests were originally designed to assess reading comprehension. Every seventh (or so) word in a passage is eliminated, and the reader has to fill in the space with the correct word (more flexible versions allow a synonym). A text is chosen that can only be understood if the topic under discussion is understood, rather like the gobbet. The omitted words are essential for making sense of the passage.

The simplest way of scoring is to count the number of acceptable words completed. You could try to assess the quality of thinking underlying each substitution, but this diminishes its main advantage, speed.

Procedures for rapid assessing

We should now look at some procedures that speed up assessment.

Self/peer-assessment
This can fractionate the teacher's assessment load in large classes, even when you use conventional assessments such as exam or assignment. Using posters, the assessment is over in one class session. But of course the criteria have to be absolutely clear, which makes it less dependable for complex, open-end responses.

If self/peer-assessments agree within a specified range, whether expressed as a qualitative grade or as a number of marks, award the higher grade. The possibility of collusion can be mitigated by spot-checking. Boud (1986) estimates that self/peer-assessment can cut the teacher's load by at least 30 per cent.

Group assessment
Carrying out a large project suggests teamwork and group assessment. Teaching large classes also suggests group assessment, but here the logic is more basic. With four

students per assessment task (whether assignment, project or whatever), you get to assess a quarter the number you would otherwise, while the students get to learn about teamwork, and assessing others, not to mention the content of what was being assessed. Considerations about allocation of assessment results apply as before (pp. 181-2).

Random assessment

Gibbs (1998) cites the Case of the Mechanical Engineer, who initially required 25 reports through the year, but as each was worth only a trivial 1 per cent, the quality was poor. He then changed the requirements: students still submitted 25 reports, but in a portfolio by the end of the semester, as a condition for sitting the final exam, but only four reports, marked at random, comprised 25 per cent of the final grade. Two huge benefits resulted: the students worked consistently throughout the term and submitted 25 good reports, and the teacher's marking load was a sixth of what it had previously been.

Feedback, open information

Make sure the students know exactly what is expected of them. Following are some things that cut time considerably.

Get assessment criteria down on a pro-forma, which is returned to the students. You don't have to keep writing basically the same comments.

Assess the work globally, but provide a quick rating along such dimensions as may be seen as desirable. You could rate them on a quantified scale, but that encourages averaging. It is better to put an X along each line, which just as clearly lets the students know where they are:

focused	◄――――――――►	unfocused
original	◄――――――――►	derivative
theoretical	◄――――――――►	atheoretical
good expression	◄――――――――►	poorly expressed
good coverage	◄――――――――►	thin
well referenced	◄――――――――►	poorly referenced
whatever	◄――――――――►	whatever not

You are letting the student know that these individual qualities are import whether or not they make a quantifiable difference to the final grade. You could do as is done in dissertations and treat them as hurdles, which have to be cleared satisfactorily before the real assessment begins.

Keep a library of comments on computer for each typical assignment you set. They can be placed in a hierarchy corresponding to the grade or performance level in which they occur. New comments can of course be added, while it saves you having to keep rewriting the common ones ('this point does not follow . . .*'*). R. G. Dromey (private communication) is developing a program that takes this much further, making assessment of lengthy papers highly reliable, feed-back rich and done in one-third the usual time.

Put multiple copies in the library of previous student assignments (anonymous, but you had nevertheless better get permission), representing all grades, and annotated with comments. Students can then see exactly what you want, that you mean it, and what the difference between different grades is (which is also likely to save time on *post mortems).*

Deadlines

Part of the felt pressure on both staff and students in large-class assessment is due to the pile-up of work, as much as to the amount of work itself. One value of multiple assessments is of course that some can be collected earlier in the seminar if the topics have been completed, but be careful not to confuse the formative and summative roles of assessment (pp. 142-3). In large classes, you have to be ruthless about deadlines. It is important to discuss your deadlines with colleagues to make sure they are evened out for the students.

Final grades and reporting assessment results

The final stage of assessing involves converting one's judgements of the student's performance into a final summative statement, in the form required by administration. This raises several issues:

1. Combining results in several assessment tasks to arrive at a final grade
2. Reporting in categories, or along a continuous scale.
3. Is there any distribution characteristic to be imposed on the results?

Combining several assessments within a unit to arrive at a final grade

As the grade awarded for a unit usually depends on performances assessed in a number of topics, and those topics will be passed at various levels of understanding, we need to decide how to combine these separate estimates to yield one final grade. Our commitment to holistic assessment makes this an important issue.

Say we have four assessment tasks: AT1, AT2, AT3 and AT4. (These could be separate tasks, or portfolio items.) Determining the final grade from these components is conventionally achieved by *weighting* important tasks so that they count more. But on what basis can you calculate that AT3 is 'worth *twice* as much' (or however much) as ATI? Expected time taken is the only logical currency I can think of, but that is more a matter of the nature of the task than of its educational value. In holistic and qualitative assessment, we must 'weight' tasks in other ways.

In selecting these tasks, presumably we wanted each to assess a particular quality. Let us say AT1 is to assess basic knowledge, the task being ideas taken throughout the course; AT2 problem-solving (a case study, group assessed); AT3 an overview of the unit (a concept map); AT4 the quality of the student's reflections on course content (a journal). Now we have a logical package, which makes a statement about what we want students to learn, and how well. The logic is that all aspects being assessed are important, and must all be passed, at *some* level of competence (otherwise why teach them?).

There are two main strategies for handling the problem of weighting and combining assessment results: working qualitatively throughout, and using numerical conversions for achieving the combinations.

Work qualitatively all the way

There are several ways of preserving your holistic purity:

1. *This dissertation model.* Pass/fail on subtasks, grading on the key task only. As long as minor tasks are satisfactory, the level of pass of the whole depends on the central task, as is the case in a dissertation. In our example, you might decide that the case study AT2 is the key task, so the qualitative grading of AT2 sets the final grade for the whole unit, as long as all the other tasks are satisfactory. If they are not, they should be redone and resubmitted (with due care about the submission and submission deadlines).
2. *The profile.* Where all tasks are of equal importance, each is graded qualitatively, then the pattern is looked at. Is the modal (most typical) response distinction? If so, the student is mostly working at distinction level, so distinction it is. In the case of an uneven profile, you might take the highest level as the student's final grade, on the grounds that the student has demonstrated this level of performance in at least one task. A student who got the same grade on all tasks would, however, see this as 'unfair'. Alternatively, you can devise a conversion: high distinction = maximum performance on all tasks; distinction = maximum on two tasks, very good on remaining ones; credit = one maximum, two very good, rest pass ….and so on.
3. *Im1ied contract.* Different tasks are tied to different grades. If students want only a pass, they do AT1 alone, say, which will show they have attended classes, done the reading and got the general drift of the main ideas dealt with. To obtain a credit, they add AT3 to AT1, showing they can hang all the ideas together. Distinction requires all for the credit plus AT4, to show in addition they have some reflective insights into how it all works. High distinction needs all the rest plus AT2, the key test of high-level functioning, the case study.
4. *Weighted profile.* Require different levels of performance in different tasks. Some require a high level of understanding (e.g. relational in SOLO terms), others might require only 'knowledge about' (multistructural), others only knowledge of terms (unistructural). All have to be passed at the specified level. This is a form of pass/fail, but the standards of pass vary for different tasks. 'Weighting' in this case is not an arbitrary juggling of numbers, but a profile determined by the structure of the curriculum objectives. The only problem is in the event of one or more fails. Logically,

you should require a resubmission until the task is passed. Practically, you might have to allow some failure, and adjust the final grade accordingly.

Convert categories into numbers

First, let us distinguish absolutely clearly between assessing the performance, which may be done qualitatively, and dealing with the results of that assessment, which may be done quantitatively. Quantifying performances that have been assessed holistically is simply an administrative device; there is no educational problem as long as it follows *after* the assessment process itself has been completed.

Quantifying can be used for two related tasks:

(a) combining results of different tasks in the same unit to obtain a final grade;
(b) combining the results of different units to obtain a year result, as for example, does the familiar grade-point average (GPA).

The GPA is the simplest way of quantifying the results of a qualitative assessment: A = 4, B = **3, C** = 2 and D = 1. You weight and combine the results as you like.

You may, however, want finer discrimination within categories. There are two issues to decide:

1 Qualitative: what *sort* of performance the student's product is.
2 Relative: how *well it* represents that sort of performance.

Issue 2 is often addressed in three levels: really excellent As (A+), solid middle-of-the-road As and As but only just (A-). Here, the original assessment of each task is first done qualitatively, then quantitatively. The final result using a four-category system is a number on a 13-point scale (A+=12 …D-=1, F=0). (Note, however, that this is not really a linear 13-point scale (12 + F), but a two dimensional structure (4 x 3 + F) that we have opened out for practical reasons.)

The results can now be combined in the usual way, but the conceptua1 difficulty is that we are back to assigning numerical weights arbitrarily: even taking an average is using a weighting system of one, which is just as arbitrary as saying that a task should be given a weighting of 2, or 5.7. Nevertheless, it is what is usually done, and it is at least convenient. When the results of different subjects have been combined, the final report can be either along the same scale, or converted to the nearest category grade. For example, if the weighted outcome score is 9.7, the nearest grade equivalent is 10, which becomes A-.

Reporting in categories or along a continuous scale

Having combined the results from several assessment tasks, we now have the job of reporting the results. This is a matter of institutional procedure, and obviously we need to fit in with that. There is no problem for us level 3 teachers where the policy is to report in categories (HD, D . . . or A, B, C…). But what if your institution requires you to report in percentages? Or, as some do, report in percentages so that they can then convert back to categories: MD = 85+, D = 75-84, credit = 65-74, pass = 50-64? (This last case is exasperating. Why not report in categories in the first place?)

All is not lost. We simply extend the principle of the 13-point scale.

1. The first step is the same. The assessment tasks are criterion-referenced; to the objectives, which tells you whether the performance is high distinction (or A) quality, distinction (or B) quality, and so on through the category system you use.
2. Allocate percentage ranges within each category according to your institutionally endorsed procedures (see Figure 9.1).
3. Locate the individual student's performance along that within-category scale.

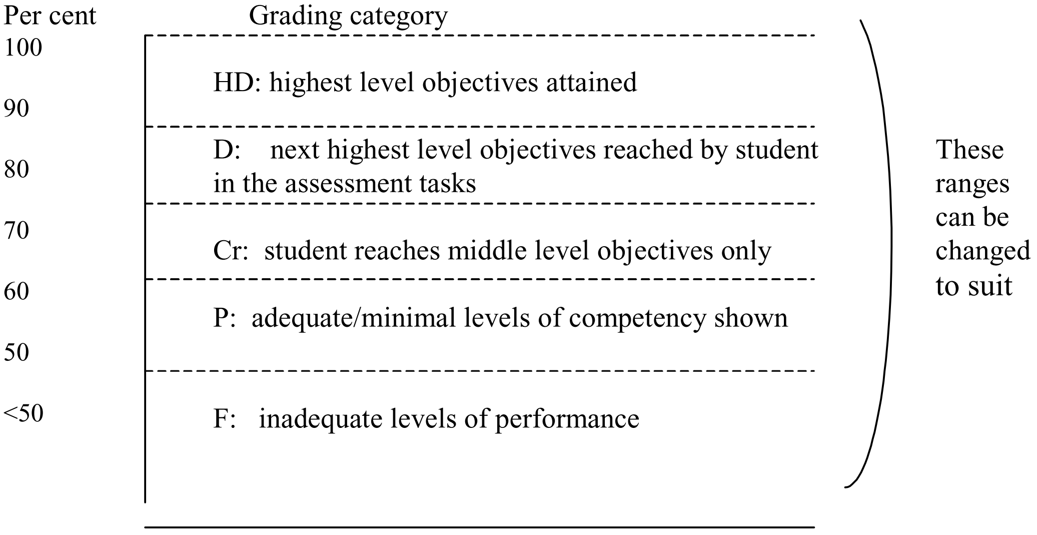

Per cent
Grading category
100
90
80
70
60
50
<50
HD: highest level objectives attained
D: next highest level objectives reached by student in the assessment tasks
Cr: student reaches middle level objectives only
P: adequate/minimal levels of competency shown
F: inadequate levels of performance
These ranges can be changed to suit

Figure 9.1: Assessing qualitatively and reporting as a percentage

Step 3 now uses a much finer scale than the previous three-level scale, something like 15 points within each category, and the student performance is quantified accordingly. You can do that by using a global or holistic rating scale, or, if you must, by awarding so many marks for this, so many for that. But at least the *major* classification into high distinction or A, or whatever system is used, has been done holistically. The rest is only a matter of fine-tuning.

Of course, this is a compromise. We have conceded defeat over the question of weighting, but the *backwash* for students is still positive. Students are likely to shoot for quality, because a category shift means a disproportionately large increase in their final score. That score also tells them something about the quality of their performance, because it falls within a range that is tied to a category. So they know the quality of their performance, and how well they did within the quality of They will also be clearer about what they would need to do to obtain a better score in future.

In sum, then, qualitative and holistic assessment can meet the logistic and administrative demands of: (a) combining assessment tasks to achieve a final grade for the unit; and (b) reporting in percentages, or any other quantitative scale, if that is what is required.

Is there any distribution characteristic to he imposed on the results?

If the answer to the above is 'yes', we cannot be so accommodating. Requiring results to fit some predetermined distribution, normal, rectangular or whatever, *cannot be justified on educational grounds.*

I am often surprised in discussing this issue at staff workshops at how many people think that CRA is pie in the sky because they *must* grade on the curve. Few institutions are in the event rigid on this point. Many 'suggest' that grades follow a distribution — 'It would usually be expected that in large classes no more than 10 per cent of high distinctions be awarded….'— but I have found that the operative word is 'usually'. In most case, it is accepted that in 'special' circumstances — for example, a criterion-referenced system — the grades of a particular class might depart from the suggested guidelines. Mind you, calling CPA 'a special circumstance' is galling, but if a special circumstance is the Trojan horse that makes aligned teaching possible, so be it.

If a teacher is employed in an institution where summative results really are required to adhere closely to some predetermined curve, there is a problem. The solution then can only be political: lobby to get the policy changed.

Implementing assessment package 2

Let us now return to the problem we faced at the beginning of the chapter implementing assessment package 2 in a class of 400 first-year students in **a** laboratory-based science course, say biology. You might remember that practicalities suggested MC as the preferred mode of assessment. We now know that there are many better alternatives. How might we now address that problem?

First, let us make the scene, a common one, more specific.

- *Class:* 400 first-year students.
- *Teaching structure:* two plenary lectures, one tutorial of 20 groups of 20 students, and one 2-3 hour lab a week, again 20 groups of 20 students. There are eight major topics introduced and variously elaborated in the lectures and tutorials over the 12-week semester.

- *Staff:* one lecturer in charge who delivers all the lectures and takes a couple of tutorials. Three teaching assistants between them take the remaining tutorials and help with the assessment. Twenty student demonstrators conduct the labs and assess the lab reports for their own groups.
- *Assessment* (existing):

		Per cent of final
Mid-semester:	1 hour MC	30
Final exam 2 hours:	1 hour MC	30
	2 essay questions	30
Lab reports		10

Institutional regulations require that at least 60 per cent of the final grade is determined by invigilated exam. The mid-semester is used to alleviate the pressure at the end of the semester, and to provide feedback to students. The MCs are all machine scored, so the main assessment load is provided by two essay questions, which are marked by checklist by the lecturer three tutors, and by spot-checking the lab reports. Say that the final occurs at the end of the examination weeks, and there is only a weekend and five working days in which to mark, collate and report the assessment result.

In previous years, there was pressure to cull the first years by about 50 per cent, in order to ease pressures in the second year, and to focus on promising research students. This pressure led to grading on a curve designed so that the bottom half received no more than a pass; credit and above thus became the *de facto* prerequisite for the second year. However, with the current realization that more students means more money, that pressure has largely disappeared, and with it the pressure for norm-referencing using predetermined grade proportions.

Problems with existing assessment

The major problem is that the assessment tasks are overwhelmingly quantitative, and address declarative knowledge. An attempt was made to offset the MCs with the essay questions, but the gesture is nullified by checklist marking. Students are not in practice encouraged to look for relating ideas, broad principles or functioning knowledge. The only non-declarative knowledge is assessed in the lab reports, but they contribute 10 per cent only and are in the event assessed by student demonstrators, not content experts. An attempt is made to provide formative feedback, apart from informal feedback in tutorials and labs, with the mid-semester, but it is in the form of marks only.

A suggested rescue package

Our present task is to design a package that would work for the given teaching structure. Let us say that resource and other limitations prevent any drastic change in the number of plenaries, labs and tutorials, and that the average assessment time per student for the final exam cannot exceed much more than 15 minutes per student (which rules out portfolios and other extended qualitative assessment tasks).

We do not immediately consult Table 9.2 under ‘rapid assessment’ and start throwing in concept maps, cloze tests, gobbets and so on. We first should specify what we *want* to assess, what our objectives are; then we might look at the most practical ways of assessing that, given the present constraints. Given the number of component assessments, the need to weight and combine them, and traditional practice, one advisable constraint would be to collate and report the assessment results quantitatively, even though we shall be using qualitative tasks for the assessments proper (see pp. 191-2).

Table 9.1: Required levels and kinds of understanding, and suitable assessment tasks

Objectives	*Kinds and levels of understanding*	*Suitable assessment tasks*
1 Basic facts, terminology	recall, recognition	MC or short answer
2 Topic knowledge	individual topics, relational, some multistructural relations between topics	gobbets, critical incidents
3 Discipline knowledge	conception of unit as a whole	letter to a friend, concept map
4 Functioning knowledge	topic or discipline knowledge put to work	problem-solving, research project
5 Laboratory skills	procedural knowledge	laboratory behaviour, lab reports
6 Monitoring and evaluation skills	metacognitive knowledge, self-directed learning	self- and peer-assessment

Table 9.1 suggests some of the levels or kinds of understanding that we should want from the students, and what kinds of assessment tasks, practical within our constraints, might be used.

1. Basic factual knowledge and terminology is suitably assessed by MC or short answer, as long as we are clear that that is all they are doing, and that these modes do not dominate the assessment package. Let us use short answer for the mid-term, which being open-ended might also show more revealing misunderstandings than an MC, and when marking time is not so pressing. MC will then be used in the final exam when time is more pressing.

2. Topics ideally should be understood at least at relational level, but ‘knowing about’ will do as long as the most important topics are understood relationally, and as functioning knowledge. The topics could then be embedded in gobbets, at the individual topic level in the mid-semester, and gobbets requiring integration of topics in the final. A critical

incident or case study in the final would also be useful; for example, the student selects a newspaper clipping of an eco-problem and relates it to topics dealt with.

3. By ‘discipline knowledge’ I mean the picture of the whole: having studied a list of topics that make up a first-year biology course, what is the student’s view of biology itself? Letter to a friend is a good way of ascertaining this (Trigwell and Prosser 1990). A description or list of topics studied (multistructural) is not good enough, a working view of an integrated subject called biology is very good (relational), a changed perspective of the biological world would be marvellous (extended abstract), if rather unlikely at this level.

4. Functioning knowledge. It is reasonable to expect that students can solve real world problems. It is suggested that six such problems are given throughout the semester as the subject of peer-assessment, much as described by Gibbs (1998), two such problems being self- and peer-assessed for inclusion in the final grade (Boud 1995).

5. Laboratory skills are mainly assessed *in situ* by student demonstrators, and probably do not go much further than the procedural level, i.e. correct performance of laboratory procedures and writing them up appropriately. Laboratory work ultimately involves functioning knowledge, but it is doubtful if it would be validly assessable in the first year under these conditions. This can be better addressed in labs in higher years.

6. Monitoring and evaluation skills, as argued elsewhere (pp. 92-3), are essential learnings for students if they are to become autonomous and self-directed in their tertiary learning, and later in their professional lives. Internalized standards of competence, which enable reflective thinking and self-direction, can be developed by self- and peer-assessment (Boud 1995; Gibbs 1998). Essentially, four of the six problems are assessed by a peer according to a marking sheet, and then each is returned to the owner. The last two problems become part of the final grade: students first self-assess on a separate sheet of paper, which is handed in, then the peer-assessment is made. If these agree within specified limits, the higher grade is taken; if they disagree, the lecturer adjudicates, and also spot-checks some of the others at random.

A range of assessment tasks has emerged here: quantitative (MC and short answer), qualitative (three gobbets, critical incident, letter to a friend, problem-solving) and procedural (lab report). For logistic reasons we need to turn all these into numbers, while retaining the qualitative nature of the majority of the assessment tasks. The qualitative tasks, with the possible exception of the problems (see below), may be assessed with SOLO, using a five-point scale within each:

SOLO level	
unistructural	
multistructural	Range
relational	1-5
extended abstract	6-10

11-15
16-20

In other words, top of the multistructural range is in conventional terms a bare pass (10 out of a possible 20) in six of the main assessment tasks. This sends a strong message to the students that 'knowing more' just will not do; you have to structure and use your knowledge.

How this applies to the problems is held in abeyance at this stage. It depends on each individual problem, but as we are also using these problems for self- and peer-assessment, the assessment procedures need to be especially clear. In short, the lecturer needs to devise a 20-point marking scheme that students can use, but there is no reason why it too shouldn't be structured along similar lines: four categories (SOLO or other), five points within each.

The SOLO scale arbitrarily but conveniently yields a maximum of 20 'marks' per task, which can be combined with the results from other tasks, including the MC. This may sound complicated but in fact it is not, as way be seen from the following assessment schedule:

Mid-semester exam	*Max. points*	*Final exam*	*Max. points*
2 gobbets, 20 each	40	1 gobbet	20
Short answer	20	1 critical incident	20
		1 letter to a friend	20
		2 problems (SA/PA)	40
		MC	20
Total	60	Total	120

With 20 points for the lab report, the total number of points becomes 200: divide by two if you want to report in 'percentages'.

The weightings here for mid-semester, final and labs are identical to those for the previous, traditional, assessment. However, these can easily be changed, if you think, say, the lab reports ought to get more and the problems less (being self-and peer-assessed); perhaps you would prefer to leave the self- and peer-assessments out of the final grade.

Let us now take a look at the marking load. Let us say that each of the qualitative assessments, problems aside, is written on no more than one page. You read this, first decide on its category (multistructural, relational) and then you rate how well it exemplifies that category on a five-point scale. This takes no more than five minutes, with practice rather less. (It will, however, be necessary for the lecturer and the TA to have a training session, and to reach a criterion of at least 90 per cent agreement allowing one category difference, which is better than the usual agreement on essay ratings using the Bloom taxonomy (Hattie and Purdie 1998).)

The time needed for assessing individual students now becomes:

For the mid-term	5 mins
The short answer test	10 mins
2 gobbets	15 mins
Total	

For the final	
MC	minimal, clerical work
3 qualitative assessments	15 mins max.
spot-checking problems	**5** mins?
Total	20 mins maximum

In addition, you will probably want to spot-check the demonstrators' making of the lab reports. If this is too much, perhaps you could cut out the critical incident, or a gobbet.

As to formative assessment, that synonym for good teaching (pp.142-3), the previous scheme did very little apart from reporting relative progress in marks. The changes suggested here for the summative assessment tasks also suggest ways in which the plenary and tutorials can be used more effectively (see Chapters 5 and 6). One would be to use the pauses in the lecture (see pp. 106-9) and the 'three-minute essay' to provide feedback: what students thought to be the main point of a particular lecture could become the focus of tutorial discussion. Like the four peer-assessed problems, also carried out in the plenaries, these essays could be required but not formally assessed before the student is allowed to sit the final exam.

We now have an assessment package that takes only a little more time, but it is manageable within the resources allowed. The assessments specifically address the higher-level objectives of the unit, so that they will encourage better quality learning from the students, will equally certainly be more interesting for both you and the students and will provide much more effective formative feedback to students.

None of these suggestions is cast in stone, however. You might prefer fewer gobbets. If we didn't operate with the restriction of 60 per cent final exam, we might have had fewer plenaries (lectures) and more out-of-class group tasks that would save assessment time. The important thing is the intention and conceptualization, not the specific techniques you use. Note that when you do rethink what you are doing to one aspect, assessment, adaptive changes occur throughout the system: objectives become clearer, teaching methods themselves improve, and of course the assessment tasks get at what they should be assessing.

Table 9.2: Some different assessment tasks and the kinds of learning assessed

Assessment mode	*Most likely kind of learning assessed*
Extended prose, essay-type	
essay exam	rote, question spotting, speed structuring
open book	as for exam, but less memory, coverage
assignment, take home	read widely, interrelate, organize, apply copy
Objective test	
multiple choice	recognition, strategy, comprehension, coverage
ordered outcome	hierarchies of understanding
Performance assessment	
practicum	skills needed in real life
seminar, presentation	communication skills
posters	concentrating on relevance, application
interviewing	responding interactively
critical incidents	reflection, application, sense of relevance
project	application, research skills
reflective journal	reflection, application, sense of relevance
case study, problems	application, professional skills
portfolio	reflection, creativity, unintended outcomes
Rapid assessments (large class)	
concept maps	coverage, relationships
Venn diagrams	relationships
three-minute essay	level of understanding, sense of relevance
gobbets	realizing the importance of significant detail
short answer	recall units of information, coverage
letter to a friend	holistic understanding, application, reflection
cloze	comprehension of main ideas

Summary and conclusions

This has been an encyclopaedic chapter. Table 9.2 is a better way of summarizing the major points on assessment tasks than section summaries.

Expressing and reporting the results of assessment

We then addressed administrative issues: how to combine results to give a single summative statement, how to report in numerical form such as percentages when assessing holistically and how to avoid grading on the curve.

When the final grade depends on performances assessed in a number of topics, passed at various levels of understanding, the different results need to be combined. Two general ways of combining results were described: consistently holistic, and doing the major assessments holistically, then converting into numbers for ease of administrative handling. The latter is a compromise but the important point is that grades are defined qualitatively in the first instance, and the result tells students something meaningful. The one problem we couldn't solve was an uncompromising insistence on reporting grades along a curve, which makes criterion-referencing impossible.

The major thrust of both chapters on assessment is really quite simple. You can't beat backwash, so join it. Students will always second guess the assessment task, and then learn what they think will meet those requirements. But if those assessment requirements mirror the curriculum, there is no problem. Students will be learning what they are supposed to be learning.

Implementing assessment package 2

Finally we returned to the difficulty facing first-year teachers in particular: how to assess qualitatively under the usual conditions of large numbers of students and poor resources.

Has this helped you with your own assessment problems? Turn to Task 9.1.

Task 9.1: Choosing appropriate modes of assessment

What key topics do you want to assess? Turn to your objectives (Chapter 3, Task 3.1):

- What less important topics do you want to assess?
- What levels of understanding of each? Use the appropriate verbs to operationalize this.
- Do the topics refer to declarative, functioning knowledge, both?
- Are there any basic facts, skills, you want to check?
- What physical constraints do you have to accommodate:
  Large-class assessment methods?
  Final exam? If so, is it invigilated?

Now choose from Table 9.3 those assessment modes that seem most suitable:

____

____

How do you propose to combine the results from each assessment task to produce a student's final grade for the unit?

Holistic throughout? ____

Holistic then convert to numbers? ______________________________

Other? ____

Comments ____

____

Further reading

Boud, D. (1995) *Enhancing Learning through Self-assessment,* London: Kogan Page

Brown S. and Knight, P. (1994) *Assessing Learners in Higher Education,* London: Kogan Page

Erwin, T. D. (1991) *Assessing Student Learning and Development,* San Fransico: Jossey-Bass.

Gibbs, G., Habeshaw, S. and Habeshaw, T. (1989) *53 Interesting Ways to Assess Your Students,* Bristol: Technical and Educational Services.

Gibbs, G., Jenkins, A. and Wisker, G. (1992) *Assessing More Students* Oxford: PCFC/Rewley Press.

Harris, D. and Bell, C. (1986) *Evaluating and Assessing for Learning,* London: Kogan Page.

Nightingale, P., Te Wiata, I., Toohey, S., Ryan, G., Hughes, C. and Magln, D. (eds) (1996) *Assessing Learning in Universities,* Kensington, NSW: Committee for the Advancement of University Teaching/Professional Development Centre, UNSW.

Stephenson, J. and Laycock, M. (1993) *Using Contracts in Higher Education,* London: Kogan Page.

The list of practical suggestions on assessment is formidable; the above is a good sample. Some are obviously one-topic: Boud on self-assessment, Stephenson on contracts. Nightingale *et al.* collate ‘best practice’ from 100 university teachers, grouped under ‘verb’ headings: thinking critically, solving problems, performing skills, reflecting, demonstrating knowledge and understanding, designing, creating, performing, communicating. The other books are good sources for ideas.

TEHE Ref: R131

Biggs, J. (1999) *Teaching for Quality Learning at University* (pp. 165-203). Buckingham, UK: SRHE and Open University Press.