

Good Assessment Practice.

Lewis Elton, University College London

L.elton@pcps.ucl.ac.uk

© Lewis Elton 2002

[Adapted from part of the University College London Good Practice Guide on Assessment, 1999, and based on L. Elton (1982), 'Assessment for Learning', in Bligh, D. (ed), 'Professionalism and Flexibility in Learning, Programme of Study into the Future of Higher Education' (The Leverhulme Study), Society for Research into Higher Education, Vol 6, pp. 106 – 135.]

Content

A. Aims, objectives and methods

- (a) Purposes of Assessment
- (b) Programme objectives
- (c) Different methods of assessment
- (d) Student motivation

B. Assessment as a measuring instrument

- (a) Standards of measurement
- (b) Reliability
- (c) Validity
- (d) The relationship of reliability and validity
- (e) The 'backwash' effect
- (f) Connoisseurship
- (g) The use of weighting
- (h) To grade or not to grade

C. Constructing examination papers.

- (a) Structuring an examination paper
- (b) When to set questions
- (c) Essay questions
- (d) Progressive questions
- (e) Multiple choice questions
- (f) Equal opportunities issues

D. Recent developments

Good Assessment Practice

Reasonably firm answers are given to some of the questions raised, but this is by no means so in all instances. It is therefore necessary to read what follows with a constructively critical mind. It is also necessary to appreciate that there may be a real tension between some of the matters advocated and the practicalities of assessment under particular circumstances, as carried out by busy academic staff.

A. Aims, objectives and methods

(a) Purposes of Assessment

Assessment can have a number of different purposes:

1. Selection and/or grading
2. Maintaining standards
3. Motivation of students
4. Feedback to students
5. Feedback to teachers
6. Preparation for life
7. Licence to practice.

There may well be others and not all of them can be achieved with a particular assessment method. Different purposes may need different methods

We believe that the overall aim of any educational programme should be that it leads to student learning. This may be obvious, but what is less obvious is whether assessment has a part to play in this. Purposes 3 and 4 help to promote learning directly, purpose 5 does so indirectly, purpose 6 does so provided one of the aims of the programme is preparation for life, but purposes 1, 2 and 7 do not in themselves encourage learning at all.

The purposes of assessment thus polarise into *assessment for learning*, also known as *formative assessment*, and *assessment for decision making*, also known as *summative assessment*. It is possible for an assessment to have both formative and summative aspects, but when the summative aspects are dominant, as they are for instance in most examination papers, formative aspects frequently get ignored. On the other hand, much course work assessment can be both formative and summative.

(b) Programme objectives

It may be obvious that what should be assessed in an educational programme is the extent to which its objectives are achieved, but in practice this is often not the case. Possible reasons for discrepancies between objectives and assessment are:

1. The programme has both process and product objectives, but only the product objectives are assessed; eg in laboratory work, the development of skills is a process objective and the production of a report a product objective, but only the latter is assessed.
2. The programme has both knowledge and skills objectives, but only knowledge objectives are assessed; eg a term essay on a particular topic also develops essay writing skills, but only the topic is assessed.
3. Some objectives are more difficult to assess than others, and only those more easily assessable are assessed (see validity and reliability below)
4. No explicit attempt has been made to link assessment to programme objectives; eg it is not possible to tell from an examination paper which objectives are being assessed.

(c) Different methods of assessment

Different assessment methods are appropriate for different assessment purposes and for the assessment of different programme objectives. Here is a list of possible assessment methods (others may be added):

- unseen examinations, structured or unstructured, with or without choice of questions
- open book examinations
- examinations with advance information about questions
- single question unseen paper
- multiple choice/ objective tests [a good guide is G Isaacs (1994), Multiple Choice Testing, HERDSA Green Guide 16]
- coursework assessment
- oral examination
- assessed report, dissertation, thesis
- assessments, the forms of which are negotiated in advance between examiners and students
- self and peer assessment
- group assessment

Some of these methods overlap, eg negotiated assessment could be in connection with an unseen examination, in which students are allowed to have a say in what is being examined..

Different methods suit different candidates better or worse, so that a mix is fairer than a single method. The introduction of course work assessment has certainly given coursework a deservedly higher profile, but it has to be accepted that its assessment is likely to be less reliable (for a discussion of reliability see below), since the work is not carried out under examination conditions. Since examination conditions are highly artificial, this feature of coursework may actually be considered an advantage, where ‘natural’ conditions are important.

Coursework assessment is usually thought to have both formative and summative properties, and this is true, but not unproblematic. If coursework assessment is preceded by wholly formative practice assessments, then students often fail to take advantage of the formative ones; if it is not, then students lack practice when they take the ‘for real’ assessment. One way of getting over this problem is to make every course work assessment count, but give two deadlines. Students can choose whether to hand in work before the first deadline, in which case it is commented on and returned for improvement, or not. Everyone’s work must be handed in before the second deadline, which is the assessment deadline, whether the work had been handed in before or not.

Quite generally, examiners might consider whether to give more choice to students, since choice - even between the frying pan and the fire - tends to motivate. Here is an example, which resulted from the first student riots at Berkeley in California. These were largely started through a demand to reduce the pressure of the ‘grade point average’ and were defused by giving students choices between being graded or simply passed. The new arrangement was that students were assessed on, say, six pieces of work or six papers, but that only three were graded, while the other three had simply to be passed. This gave students the opportunity to shine at what they thought they were better at and into which they had put more effort.

Coursework assessment is particularly suitable for assessing process objectives, and essential for assessing true creativity and genuine problem solving abilities, neither of which can normally be assessed under the stresses and time pressures inevitable in formal examinations.

Oral examinations appear to be used for three very different purposes:

to test verbal skills and the ability to argue a point
to verify that students’ written work in, say a dissertation, is their own
to help to make decisions on border line cases.

The first two purposes are legitimate, the third is more doubtful, since a brief and stressful oral, often conducted by an external examiner whom the student has not met before, is unlikely to lead to reliable decision making. **What is essential for all oral examinations is that examiners must receive training in interview and interpersonal skills.**

Negotiated, self and peer assessment moves assessment from wholly teacher determined to a joint enterprise between teachers and students. It is thus particularly appropriate for student centred programmes, but both teachers and students have to develop appropriate skills and attitudes before such forms of assessment are used.

Group assessment presents two particular problems:

How to assess the work of the individuals in the group, if not all contributed equally to the product which is to be assessed

How to assess the group process.

The first problem can be tackled in several ways:

Students are given equal marks, irrespective of their contributions

The individual student contributions are elicited through individual vivas (not really recommended) Students are given a total mark and asked to divide it appropriately between themselves (in this and the following method it is highly desirable to negotiate appropriate criteria in advance of the work which is to be assessed) Self and peer assessment are used at least in part.

The assessment of the group process could in principle be done through teacher observation, but this tends to disrupt the process. A much better way is to rely on the peer assessment of the participants, in which case it is essential to negotiate appropriate criteria in advance of the work which is to be assessed.

(d) Student motivation

Thirty years ago the commonly held view was that student should be motivated by their love for a subject and that examinations were a necessary evil, which detracted from this desired motivation. It is likely that few students ever held this view and that few teachers held it when they were students. More recently, it has been generally accepted that - in the words of an American author - 'grades are the campus currency', ie just as people at work are motivated by earning money, so students are motivated by earning marks. However, while people in work would not work without being paid, their motivation is often the interest of the work. Similarly, **it has been found that once students are satisfied that they are being fairly prepared for their examinations, their motivation is governed more often than not by the intrinsic interest of their study.** Thus money and marks act as a trigger - without them there is no intrinsic motivation, but once the trigger operates, motivation depends much more on intrinsic interest than on size of the reward, whether in money or marks.

When the objectives being assessed diverge from the declared learning objectives - and this is distressingly common - students find themselves in a classical double bind situation. They either please their teachers through their intrinsic interest and fail their examinations or they concentrate on passing the examinations and displease their teachers. Sensible students choose the latter course, those that choose the former may well eventually take their exams from their hospital beds.

B. Assessment as a measuring instrument

Any measuring instrument involves a comparison with a *standard*, and aims to be *reliable* and *valid*, concepts which will be explained below in connection with assessment. In addition, any measurement is subject to the *Heisenberg Uncertainty Principle*, according to which the act of measurement changes what is measured in a not wholly predictable manner.

(a) Standards of measurement

In scientific measurement, a measurement is a comparison between the result of the measurement and an absolute standard, eg one compares one's weight as read by a weighing machine with the standard kilogramme in the international Bureau of Weights of Measures in Paris. In educational measurement, we assess students either against some pre-specified performance criteria or against the performance of a representative group of comparable students. The former, which is called *criterion referenced*, suffers from the same problem as the specification of objectives; ie it can suffer from both over and under specification. It is rarely used in university examining, except in certain professional examination, such as pharmacy. The latter, which is called *norm referenced*, depends on a sufficiently large and representative group of students being available. This may be the case at A-level, although even there the norms are not the sole criterion, but is effectively never the case in universities. The normal practice in universities would appear to be a mixture of criterion and norm reference. The borders between classes are permanently fixed, eg a First is above 70%, which implies that this mark satisfies the criteria for a First. This conclusion is however modified by current groups of students being compared for their quality with those of previous years, and the proportion in each degree class then being adjusted accordingly. This procedure, which can hardly be described as precise, relies considerably on standards from year to year (it may be noted that 'standard' is used here in a very different sense from that used earlier in this section), maintained through the memories of the examiners. While this does result in a similar degree class distribution for a particular degree

examination in a particular university from year to year, it can and does lead to very different degree class distributions for a particular degree examination in different but comparable universities.

The belief that it is possible to fix the same class boundaries numerically also for different subjects is, however, wholly invalid, if marking practices differ substantially between different disciplines. It has been known for a very long time that mathematical type marking usually stretches over the whole scale 0 - 100%, while essay type marking more typically ranges from 30 to 70%. The argument for this difference, that it is possibly to reach perfection or total failure in mathematics but not in essays is absurd; both in mathematical and essay type assessment, perfection and total failure ought to be defined in terms of what is reasonable for the examination in question. Present indefensible marking practice is responsible for both the lack of Firsts in subjects like sociology and history, and the higher proportion of failures in subjects like physics and engineering. It also leads to absurdities in modular and multi subject degrees where it is essential for class distributions to be comparable in different subjects. Fortunately, most modular degree programmes now have marking schemes specified in terms of comparable achievements, so that the distribution of marks for different subjects is similar.

Finally, it is important to remember that class boundaries were settled a long time ago when virtually all assessment was based on finals papers. To take them over into schemes where a substantial part of the assessment is by other means is quite indefensible (rather like keeping the marks on a thermometer the same, but change from mercury to alcohol). In particular, it is well known that on average course work assessment leads to higher average marks and smaller spreads of marks. It is very likely that the apparent grade inflation over the past twenty years is due largely to examiners' ignorance of simple principles of measurement and is not a reflection of either improved learning or greater lenience in marking.

(b) Reliability

There are two kinds of reliability, which a measuring instrument must satisfy:

- 1. Two people who use the same instrument to measure the same thing should get the same result (examiner reliability)**
- 2. Two supposedly equivalent instruments should give the same result when measuring the same thing (test reliability)**

Innumerable investigations have shown that educational assessment at best is only moderately reliable in the first sense. Test reliability is far more difficult to verify, since it requires two supposedly equivalent tests to be given under the same conditions to the same students. However, it has been investigated, with the result that by and large the best students came out best on both tests and the worst students worst on both, but the ranking of middling students was very different in the two tests investigated. This is particularly important for multiple choice tests, which are 100% reliable in the first sense, but can be very unreliable in the second.

(c) Validity

An assessment is valid, if it assesses what it is intended to assess, which is usually specified in terms of the learning objectives which are to be assessed. Since the performance which is assessed through any form of assessment is inevitably different from the corresponding performance under more normal circumstances, the validity of the assessment can only be gauged by appropriate experts and can never be 100%. Points which experts will look for are that the assessment must fairly reflect the programme objectives which are to be assessed, ie it must not be testing

- outside the programme objectives
- too selectively within the programme objectives
- at an inappropriate level of the programme objectives.

One way to apparently ensure greater validity is to specify objectives very precisely, in the extreme in behavioural terms. However, this distorts the learning that is being assessed, since it focusses it in a most constraining way. Good learning is always more than being able to jump through pre-specified hoops,

however well the hoops are pre-specified. **It is generally agreed that learning objectives must not be specified either too tightly or too loosely, if the learning is to be validly assessed.**

The validity discussed so far is called '*face validity*', because it involves the assessment being faced by an expert. There are other kinds of validity, which are concerned with the purposes of assessment, such as *predictive validity* in connection with future performance and *teacher feedback validity* in connection with improved teacher performance. In contrast to face validity, these other forms of validity are testable.

(d) The relationship of reliability and validity

Unfortunately, reliability and validity are not mutually independent from each other. Programme objectives which can be tested most reliably have two outstanding features:

- they largely test memory, since this leads to greater agreement between examiners than the assessment of higher learning objectives, eg any kind of understanding
- they treat all candidates equally.

Unfortunately, these are the very objectives which are of comparatively little importance in degree programmes, where it is usual to expect higher abilities and skills to be developed, and where increasingly learning programmes differ for different students. **Thus there has to be a trade off between reliability and validity.** The alternative, ie to test for memory and to treat everyone the same in the assessment, even though that is not what the programme objectives demand, is not however unknown. Its effect will be discussed in the next section.

(e) The 'backwash' effect

It has already been indicated that the Heisenberg Uncertainty Principle, according to which measurement changes what is measured is of great importance in assessment. It is even more important here than in physics, because when a measurement affects people, it affects their behaviour also before the measurement takes place. This is the 'backwash' effect, according to which **students' learning is guided by the assessment to come and the objectives being assessed become the students' learning objectives.** In particular, if assessment assesses memory learning, then students tend to engage in memory learning, whatever their teachers may say. (This point was made already under 'student motivation'.)

The consequences of the backwash effect lend strength to the argument that high validity is more important than high reliability, although this is no excuse, once high validity is assured, for not doing one's utmost to increase also reliability. However, the result is almost inevitably a less 'fair' assessment, because fairness appears to demand that everyone is treated the same and is assessed in a highly verifiable manner. Perhaps the biggest difference between the educational and the so-called real world is this insistence on fairness in the former as a criterion above all others in assessment. Once it is appreciated that life is not fair and that education is a preparation for life, it will perhaps become more acceptable to take fairness off its high pedestal in the interest of better and more relevant education.

(f) Connoisseurship

The unreliability of assessment is particularly pronounced in disciplines with a strong creative and/or aesthetic component, but if it is accepted that all good assessment is to some degree unreliable, as regards both standards and individual performance, then perhaps it may become acceptable to use the concept of connoisseurship also in more usual disciplines. Connoisseurs are persons who, through training and experience, can make expert and reliable judgments in their specialist fields, which is not a bad definition of a good external examiner. Their judgment is accepted, because it is seen as both expert and reliable, and should therefore provide appropriate assessment standards. However, external examiners should normally have explicit checklists, so that their judgments are not totally holistic. A better model for the examiner may therefore be the judge in a sporting competition, such as ice skating or gymnastics, where both examiners and examinees know the dimensions on which the judgments will be made.

(g) The use of weighting

It is not uncommon to find that forms of assessment that are inherently of low reliability are given a small weight in the overall assessment, so as to reduce the effect of the unreliability. This is thoroughly bad practice, especially as very often these less reliable assessments are necessarily used to assess the more important learning objectives. Weighting should reflect the importance given to learning objectives, not the reliability with which they can be assessed.

(h) To grade or not to grade

Increasingly, the grading of degree work along a single dimension is being called into question. Is it really meaningful to describe three or four years' work in terms of a single number? If not, then it is time that the reporting of degree results were done in terms of a profile, in which some assessments would be on the traditional classified basis, some would be pass fail, some would be brief reports in words and some might be no more than a certification of attendance. Only such variety could meet the needs of different learning objectives in the years to come. It is important to realise that the American 'transcript' would be little more than a cosmetic improvement. Admittedly, it would give more information than a single class does, but it would still be in terms of the same kind of classification, whatever the learning that is being assessed. And as it is all too easy to calculate from it the 'grade point average', we might well be back with the single dimension in the end. Incidentally, reporting in terms of a profile also gets over the problem of different weights given to different assessments.

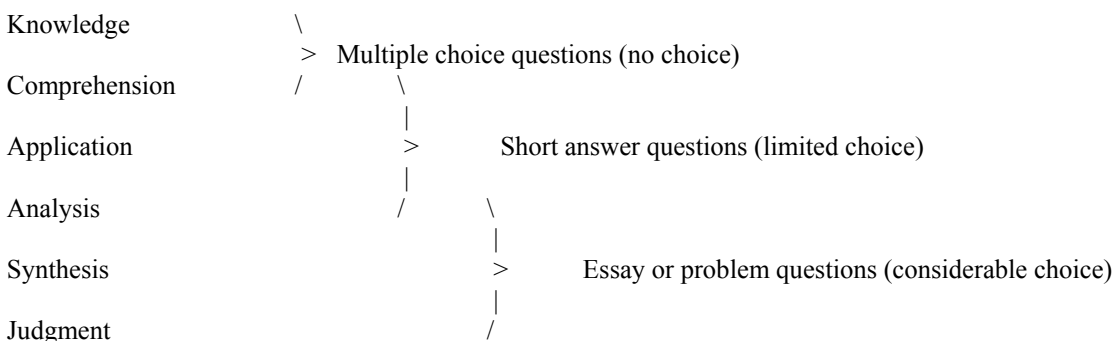
C. Constructing examination papers.

(a) Structuring an examination paper

The practice of unstructured examination papers, in which students are asked to select a portion of the questions set, which are ostensibly of the same difficulty, is only sound if the knowledge being tested is of very minor importance and if the academic skills tested by the different questions are of comparable difficulty. A broadly speaking hierarchical scale of such skills (developed from the work of B S Bloom) might be as follows:

- Basic knowledge, ie little more than memory work
- Comprehension, eg simply rearranged knowledge
- Application of knowledge, eg to practical problems
- Analysis, eg relation to basic principles
- Synthesis, eg bringing together relevant disparate considerations
- Evaluation, ie judgmental treatment of knowledge, analysis and synthesis.

Clearly, **basic knowledge should be treated comprehensively and not selectively, while the higher skills, which are concerned with academic skills applicable to the basic knowledge, can be exhibited with any basic knowledge. It is therefore reasonable that students should have a choice of knowledge base in which they demonstrate their academic skills.** This leads to the idea of a structured examination paper, with progressively more choice (or of course to different types of papers for the different levels of the hierarchy). A structured paper might look as follows:



(b) When to set questions

Many academics set questions only after they have completed a course for their students, partly so that the choice of questions will not influence their teaching and partly because they may not be certain until the end as to what they may not have time to teach and therefore should not be in the examination. The first consideration is laudable, but misguided, there is likely still to be a close relationship between teaching and questions set. The second consideration works on the not to be encouraged assumption that only that can be expected to be learned which has been explicitly taught, usually in lectures, rather than also that acquired from eg guided reading.

Others set questions well before they start their course, so as to distance them from their teaching. Unfortunately, this is likely to lead to stereotype questions, ie the very obvious ones which one is likely to think about when one is not close to the teaching.

The third way is to set questions as one proceeds with a course. This procedure is most likely to lead to the best questions and, provided one is aware of the dangers in teaching to the questions and avoids them, almost certainly the best.

Whichever method is used, it is desirable to set rather more questions than are needed and then to construct an examination paper which balances the various requirements, in particular the balance between the different items in the skills hierarchy.

(c) Essay questions

Particular care must be taken with essay questions, which usually contain an operational word, such as 'discuss', 'compare', etc.[For a more thorough treatment, see L Hamp-Lyons and B Heasley (1987), 'Study Writing', Cambridge University Press, pp. 140 – 142.] Such words should have definite meanings, which are shared through previous preparation by teacher/examiners and students. Unfortunately this is rarely the case, the understanding being generally tacit for the teacher or examiner and for that reason not shared with the students. There is no way of knowing whether this tacit understanding is the same for all examiners, and so only those students who happen to share the tacit understanding of a particular examiner are likely to do well. What is needed for a fair assessment (here fairness is both possible and desirable) is for that understanding to become explicit between teachers and students and for examiners to agree to it. Possible operational words are:

- define
- describe (also: list)
- analyse
- explain (also: account for)
- assess
- compare
- contrast
- argue (also: justify)
- critically examine/evaluate

and some of these are explained more fully in the attached article. They have different meanings and they correspond to different levels in the hierarchy of academic skills. All this must be taken into consideration, when constructing a balanced examination paper.

The perhaps most common operational word, ie 'discuss', is not in the list, as it has too many different meanings in different circumstances, each one of which is better expressed by one of the words in the above list. Could it be that the popularity of this word in examination questions reflects the not uncommon attitude in examiners to deliberately be vague in a question, in order to see how students tackle it and then match their responses to the examiner's preconceived perceptions, which are of course unknown to the examinees? [Note that 'Study Writing' does allow 'discuss', but calls it 'one of the most difficult types of essay question'.]

There is also the essay question, which is literally a question. Here are three such questions, which all were alternative questions in a genuine paper:

Who were the “New Karaites”?

How has the excommunication of Spinoza been explained?

How significant was the Thirty Years War for the Jews in Western and Central Europe?

Without any familiarity with the subject, it would appear that the first calls for a purely factual answer, while the second expects an evaluation of different possible explanations, and the third could legitimately be answered with the single word: “very”. This is an indication that **questions which are literally questions are in general unacceptable.**

Finally, it is worth raising the issue of using computers in essay examinations. Students are now strongly encouraged to word process their course work. Is it then right that they should return to pen and paper for their examinations? And if not, can the problems associated with bringing computers into examinations be overcome? There are at present no answers to these questions, but they are worth asking.

(d) Progressive questions

Questions which are in several parts, and in which information in the earlier parts helps in responding to the later ones, are common in the sciences. In such composite questions, the first part is often a purely memory question, the second part uses the information contained in the first part in solving a problem and the third part may similarly use it for a more theoretical question. In such cases, the wording of the first part gives information on the later parts, eg what might be a good starting point for solving a particular problem, and thereby makes the second part easier than it would have been, if the first had been omitted. Since a most important aspect of problem solving lies in the identification of a suitable starting point, this may devalue the question as a test of problem solving skills. Furthermore, if such a question is now set in an open book examination and the first part is omitted, as being purely memory work and therefore unsuitable for an open book question, then the problem part may have been made considerably harder by this omission, a fact that is often not appreciated.

(e) Multiple choice questions

The perception that MCQs can test only for knowledge recall is doubly wrong: all MCQs test for recognition rather than recall, but they can test at all levels of the hierarchy discussed in section C(a). However, the higher the level to be tested, the more difficult is it to set good questions. Quite generally, **just because MCQs give no information concerning students’ thought processes, it is particularly important that they should be of a high professional standard.** Examiners should not use MCQs, unless they have been trained in their use, and this is true even if the MCQs are taken from professionally generated question banks.

(f) Equal opportunities issues

This may be a good point to refer to equal opportunities issues. The following points ought to be considered in connection with any questions set in an examination:

- Is there a race or gender bias in the question?
- Have the needs of disabled students been adequately catered for?
- Is the language chosen such as not to unduly handicap those for whom English is not their first language?
- Has some inappropriate cultural bias been introduced?
- Has the anonymity of scripts been assured?

D. Recent developments

The matters discussed so far, with the possible exception of self and peer assessment, all relate to well established practices. **Recent developments have concentrated on forms of assessment, which take account of the movement to use assessment in order to encourage active and student centred learning, which make it difficult to assess all students in the same manner, as well as on the necessity in these days to make assessment more cost effective.** An excellent and easily readable book, which brings traditional concerns up to date, is S Brown and P Knight, *'Assessing Learning in Higher Education'*, Kogan Page 1994. This should be read in conjunction with S Brown and B Smith, *'Getting to Grips with Assessment'*. SEDA Special No 3, Staff and Educational Development Association, 1997, which gives practical advice and check lists. Both are strongly recommended, as is the booklet *'Strategies for diversifying assessment in higher education'* by S Brown, C Rust and G Gibbs, The Oxford Centre for Staff Development, Oxford 1994.