

Originally used as a template for a module on curriculum design, these files are intended to be used and adapted as required (as long as acknowledgement is featured where material is reproduced). Everyone will have their own style of presenting and organising material, so the outlines provided below are presentation guidelines only.

Accompanying notes to the lecturer's resource

Assessment presentation

PRESENTATION 1: Introduction to assessment

The purpose of these presentations are to introduce assessment so that students will be able to plan, implement and evaluate different forms of assessment, and to address the notion of 'fitness for purpose': the kind of assessment that one undertakes should be appropriate for the purposes that one has in mind. Different kinds of assessment suit different purposes, and the module considers different purposes and uses of assessment.

Assessment is not simply testing; it is much broader than that. The module will look at a range of purposes, kinds and uses of assessment. It will examine some of the latest developments in assessment, and will provide students with the opportunity to plan and devise different kinds of assessment. Much assessment in schools is simply testing, and the strengths and weaknesses of this will be explored. The module examines performance assessment and authentic assessment. If we are to make the most of assessment then it needs to break away from simply testing, and it must become much more formative and to assess higher order thinking. The implications of formative assessment and the assessment of higher order thinking are examined. Finally, as in all assessments, reliability and validity are key issues to be addressed, and the module closes by examining these two issues.

The module uses a combination of lecture input with workshop activities. The assessment, as with the previous modules, is on a group basis, and this will take place in the final session, through a presentation by each group, and through the supporting documentation provided. It comprises a task that will be prepared during several of the sessions of the module.

The assessment of the module, as will be discussed later in the module, is by an assessed presentation and supporting handout material, to be done on a group basis.

Slide 1

Slide 2

© Keith Morrison, 2004

Published on the companion web resource for *A Guide to Teaching Practice* (RoutledgeFalmer).

Slide 3

In many countries there are several different systems of assessment, and this slide identifies some of them. Different systems of assessment may use different kinds of assessment, for example diagnostic tests, standardised tests and examinations may use largely written forms of assessment, whilst vocational and occupational assessments may use more practical forms of assessment. Teacher assessments may use written forms, or course work, or a range of classroom activities. Examinations tend to be summative (end of course) whilst other forms can be before or during the course, for example for diagnosis or ongoing feedback to the student.

Slides 4 and 5

These two slides build on work that was published in 2002 in the United Kingdom and the USA. Whether or not we wish to use the assessments for educational purposes alone, there are several perhaps unintended consequences of assessment. We can see from these slides that many of the consequences here tend to be negative rather than positive.

Slide 6

There is a range of *purposes* of assessment. These can be divided into primary purposes (the main purposes) and secondary purposes (the lesser purposes).

Assessment serves a series of *primary functions*, being used for:

- *certification*, qualifying students for their lives beyond school by awarding passes, fails, grades and marks;
- *diagnosis*, identifying a student's particular strengths, weaknesses, difficulties and needs in order that an appropriate curriculum can be planned;
- *improvement of learning and teaching*, providing feedback to students and teachers respectively so that action can be planned, moving away from marks and grades and towards comments, discussions, and suggestions for how students and teachers can improve – a formative intention. It also enables greater precision in matching to be addressed;
- *to select* for future education, setting and banding, options, level of examination entry;
- *to provide evidence of achievement* of curriculum success;
- *to see the extent* to which intended learning outcomes have become *actual* learning outcomes;
- *to chart rates of progress* in learning;
- *to compare students*, for example with others in the class, set, year, school or indeed with national levels of performance;
- *to report* that which students can do and have achieved.

Ask the students which of these several primary purposes are served more than others in the UK, and why that might be.

Slide 7

Assessments can serve a series of *secondary functions*, being used for:

- *accountability* of teachers and students to interested parties – to report on standards;
- *evaluation* of the quality of teaching, learning, curricula, teachers, schools and providers of education;
- *motivating students and teachers*, though this is dependent upon the type of assessments adopted – tests tending to be demotivating whilst formative assessment and feedback tending to be more motivating.
- *discipline*, though it is questionable whether it is acceptable to lower or raise grades gained from assessments dependent on students' behaviour or misbehaviour;
- the *control of the curriculum*, for the 'backwash effect' on the curriculum is strong in 'high stakes' – external – assessment. High stakes assessment, where much depends on the results of the assessment (e.g. university entrance, employment prospects, graduation), is characterised by examinations rather than more informal methods or methods which use much teacher assessment.

It is important to be clear on one's purposes in assessment, for, as will be argued later, the choice of method of assessment, follow-up to assessment, types of data sought, types of assessment are all governed by the notion of *fitness for purpose*. Several of the purposes set out above are in a relation of tension to each other. For example using assessments for the purposes of selection and certification might be intensely demotivating for many students and may prevent them from improving; the award of a grade or mark has very limited formative potential, even though it would be politically attractive; internally conducted assessment has greater educational validity than externally conducted assessment. Using a diagnostic form of assessment is very different in purpose, detail, contents and implementation from assessment by an end-of-course examination. Using assessment results as performance indicators can detract from improvement and providing formative feedback to improve learning. The notion of *fitness for purpose* returns us to a central principle, *viz.* the need to clarify and address the objectives of the exercise. We support the view that student teachers should be concerned with diagnostic and formative assessments that are steered to improvements in teaching and learning, as these are more educationally worthwhile and practicable over the period of a teaching practice. The purposes of assessment here are educative rather than political or managerial.

Assessment can have a *backwash* effect, for example to influence the contents and pedagogy of schools leading up to public examinations, and a *forward* effect to support learning aims.

Activity 1: 35–40 minutes

In small groups, where they are sitting, ask the students which of these several secondary purposes are served more than others in UK, and why that might be. Ask them to

consider the strengths and weaknesses of using assessment for these secondary purposes, i.e. to ascertain how acceptable it is to use assessment for these purposes. Give the student no more than fifteen minutes to prepare a series of responses to the strengths and weaknesses here, and then have no more than a twenty minute feedback session, putting their responses on the whiteboard in two columns: strengths and weaknesses. The students do not have to hand in anything for this activity; it is simply a sensitising activity.

Break

Slide 8

There are several *types* of assessment, for example:

- Norm-referenced assessment
- Criterion-referenced assessment
- Domain-referenced assessment
- Diagnostic assessment
- Formative assessment
- Summative assessment
- Ipsative assessment
- Authentic assessment
- Performance assessment.

(Maybe have these written on the whiteboard in advance)

The module will address each of these, though it will take several sessions! Different types of assessment serve different purposes of education and assessment, and different types of assessment require different kinds of information and assessment evidence or data.

Slide 9: Norm-referenced assessment

A main type of assessment is *norm-referenced* assessment. A norm-referenced assessment measures a student's achievements compared to other students, for example a commercially produced intelligence test or national test of reading ability that has been standardised so that, for instance, we can understand that a score of 100 is of a notional 'average' student and that a score of 120 describes a student who is notionally above average. The concept of 'average' only makes sense when it is derived from or used for a comparison of students. A norm-referenced assessment enables the teacher to put students in a rank order of achievement. That is both its greatest strength and its greatest weakness. Whilst it enables comparisons of students to be made it can risk negative labelling and the operation of the self-fulfilling prophecy.

In norm-referenced assessments there are two main groups to which comparisons are made. If the test is standardised to the wider population (e.g. a national test or public examination of the whole population of 16-year-olds) then the individual's result can be placed at a point relative to the national norm. In many teacher-devised assessment it is the group or cohort than is the group to which reference is made (e.g. a class), in which case comparisons can only be made to those in the class in question.

Slide 10

In a norm-referenced test at the end of a course it may be decided in advance that 5 per cent of the students will gain grade A, 20 per cent will gain grade B, 40 per cent will gain grade C, 20 per cent will gain grade D, 10 per cent will gain grade E and 5 per cent will fail, as on the slide. One can see implicit in this the bell-shaped curve of distribution of abilities. This guarantees that a certain percentage of students will gain certain grades, almost regardless of their *absolute* ability: the grades are *relative* rather than absolute, i.e. they do not *necessarily* denote outstandingly good or outstandingly bad performance; they may, but it is not guaranteed.

Now, let us say that the test was conducted on two successive years. In the first year the group was generally very bright; the required percentages were placed in the various grade groups as required. When the test was taken in the second year the group was generally very poor; nevertheless the requirements for percentages receiving particular grades was the same, so the same percentage of the poor year achieved an A grade as for the bright year, i.e. the grade A meant something different for each year. We can have 'a good year' and 'a bad year' of students, even though the grades awarded adhere to the same distributions. This may be unfair, as good students in a year in which there are many good students may not do as well as good students in a year in which there are fewer bright students. Indeed, if we have two schools, one of which is very poor and the other which is very good, it might behove the poor school to adopt norm-referencing as it *guarantees* a proportion of grade As, Bs, Cs, Ds, Es, and a tiny number of failing students regardless of their actual ability, and this could put that school in a favourable light.

Just as a norm-referenced system guarantees a certain proportion of high grades, e.g. A and B, so, by definition, it also guarantees a proportion of low grades and failures, regardless of actual performance. The educational defensibility or desirability of this may be questionable: a 'good' student may end up failing or scoring poorly if the class or group of student with whom she/he is being compared is even better. Norm-referencing may be useful for selection, but it may not be equitable.

Norm-referenced assessments, because at heart they are used to compare students, can lead to competitive environments.

Slide 11: Criterion-referenced assessment

Criterion-referenced assessment was brought into the educational arena by Glaser in 1963. Here the specific criteria for success are set out in advance and students are

assessed on the extent to which they have achieved them, without any reference being made to the achievements of other students (which is norm-referencing). There are minimum competency cut-off levels, below which students are deemed not to have achieved the criteria, and above which different grades or levels can be awarded for the achievement of criteria – for example a grade A, B, C etc. for a criterion-referenced piece of course work. A criterion-referenced test does not compare student with student but, rather, requires the student to fulfil a given set of criteria, a predefined and absolute standard or outcome.

Criterion-referenced assessment specifies clearly what has to be known, shown, learned, and done to meet the criterion, in very concrete and detailed terminology. It relates directly to what has been taught and learned in a programme, e.g. the ability to apply a mathematical formula, the ability to use the periodic table, the ability to add three single digits.

In a criterion-referenced assessment, unlike in a norm-referenced assessment, there are no ceilings to the numbers of students who might be awarded a particular grade; there are no maximum numbers of proportions for each grade. Whereas in a norm-referenced system there might be only a small percentage who are able to achieve a grade A because of the imposed quota (the ‘norming’ of the test), in a criterion-referenced assessment, if everyone meets the criterion for a grade A then everyone is awarded a grade A, and if everyone should fail then everyone fails, i.e. the determination of grades is not on a quota system but is dependent on the achievement of the criteria themselves, *regardless* of how many others have or have not passed the test. If the student meets the criteria, then he or she passes the examination.

A common example of a simple criterion-referenced examination is a driving test. Here a dichotomous view of criterion-referencing is adopted wherein the driver either can or cannot perform the elements of the test – reversing round a corner without mounting the pavement, performing an emergency stop without ejecting the examiner through the windscreen, overtaking without killing passers-by. The example of a driving test is useful for it indicates a problem in criterion-referenced assessment, *viz.* whether a student needs to pass each and every element (to meet every single criterion) or whether high success in one element can compensate for failure in another. Strictly speaking in an overall criterion-referenced system the former should apply, which is draconian since a student could fail a whole course by dint of failing one small element, a common feature of the driving test. This is a problem that has been recognised in modular courses, particularly where one or more compulsory modules have to be passed. This also raises another problem in the educational sphere, *viz.* that the dichotomous view of pass/fail – *either* able *or* unable to perform a task, is unhelpful to teaching and assessment. Abilities are not of the ‘either/or’ type where one can or cannot do a task; rather, a task can be performed with differential degrees of success in its different elements or as a whole.

Another example of a criterion-referenced assessment is the music performance examinations of the Associated Board of the Royal Colleges of Music. Let us say that I am taking one of their piano examinations that requires me to play the scale of C sharp

minor in double thirds for four octaves. If I play it successfully I pass that element of the examination; if I play it very well I receive a higher mark in that element of the examination; if I play it to the standard of a concert pianist I receive an even higher mark. Then I fail my performance of a Bach fugue, missing notes, failing to bring out the fugue's subject in the stretto sections and playing the wrong notes, thereby failing this element of the examination. However, the music examination differs from the driving test in that my marks in the former are *aggregated* so that one strong element can compensate for one weak element, so that, overall I might pass the music examination even though I am carrying a failed element.

Slide 12: Advantages of using marks

The example of the piano playing examination is useful also, for it indicates that a measure of subtlety can be used in handling marks. There are several advantages in using *marks* rather than *levels* for assessment purposes:

- 1 it enables partial completion of a task to be recognised, wherein students can gain marks in proportion to how much of the task they have completed successfully;
- 2 it affords the student the opportunity to compensate for doing badly in some elements of a task by doing well in other elements of a task;
- 3 it enables credit to be given for successful completion of parts of a task, reflecting considerations of the length of the task, the time required to complete it, and the perceived difficulty of the task;
- 4 it enables weightings to be given to different elements of the task and to be made explicit to the students;
- 5 scores can be aggregated and converted into grades.

Slide 13: Problems of aggregation

The question of aggregation is troublesome for criterion-referenced tests. If one were being true to the specificity of criterion-referencing then this would argue against aggregation at all, as aggregation – by collapsing details in an overall aggregate – loses the very specificity and diagnostic/formative potential upon which criterion-referencing is premised. However if one fails one criterion out of a large number of criteria in an assessment then one fails the overall assessment.

The problem of aggregation in criterion referencing is compounded because criterion-referencing suggests the compilation of long lists of criteria for each element of an assessment, echoing the dangers of behavioural objectives (addressed in module one). The parallel with behavioural objectives is not idle, for, just as with behavioural objectives, the danger of criterion-referencing is that one only plans for and assesses the observable, performatory and often superficial, trivial aspects of education and proceeds to record them in pages of tick boxes. Attempts to minimise these problems and to make criterion-referenced tasks manageable can easily result in generalised and inaccurate data that are meaningless.

Further, if aggregation is to occur, one has to be clear what exactly is being aggregated. For example one could be aggregating components that were so dissimilar to each other, in terms of levels of difficulty, as to yield a meaningless composite. This can be illustrated by an example in English. Let us say that I receive a grade 3 for spelling and a grade 5 for reading; if I aggregate the scores to give me an average grade I will be awarded a grade 4. But what does this grade 4 really mean; what does it indicate? It is a meaningless number that tells me nothing about my specific achievements in the contributing criteria; I am unable to use the result for subsequent planning – the formative potential of criterion-referencing has been lost. What if there was no parity in actual difficulty between a grade 3 in spelling and a grade 3 in reading (which is likely to be the case)?

The task, then, for users of criterion-referenced assessments is to select limited *sample* criteria that fairly represent a wider field, balancing specificity and generality whilst adhering to notions of: (i) *item discriminability* – the potential of the item concerned to be answered correctly by those students who have a lot of the particular quality that the item is designed to measure, and the potential of the same item to be answered incorrectly by those students who have less of the particular quality that the same item is designed to measure; (ii) *item difficulty* – where the item is not so difficult that nobody answers correctly and not so easy that everyone answers it correctly.

Norm-referenced and criterion-referenced tests at first sight appear to be mutually exclusive and, indeed, in many contexts they are. However we should not overlook the fact that implicit in criteria are norms of what we might expect a student achieving a grade A to be able to do, and how this is different from a student achieving a grade B, and so on. Further, using criterion-referenced tests still enables students' performance to be compared, e.g. comparing the grades of groups of students in a class, school, local education authority and so on.

In devising assessment, teachers will need to be clear on whether they are going to operate a norm-referenced system of grades and marks (with quotas for each grade or mark range) or a criterion-referenced system, with everyone being able to score highly.

A criterion-referenced test provides the researcher with information about exactly what a student has learned, what she can do, whereas a norm-referenced test can only provide the researcher with information on how well one student has achieved in comparison to another, enabling rank orderings of performance and achievement to be constructed. Hence a major feature of the norm-referenced test is its ability to discriminate between students and their achievements – a well constructed norm-referenced test enables differences in achievement to be measured acutely, i.e. to provide variability or a great range of scores. For a criterion-referenced test this is less of a problem, the intention here is to indicate whether students have achieved a set of given criteria, regardless of how many others might or might not have achieved them, hence variability or range is less important here.

The question of the politics in the use of data from criterion-referenced examination results arises when such data are used in a norm-referenced way to compare student with student, school with school, local authority with local authority, region with region (as has been done in the United Kingdom with the publication of ‘league tables’ of local authorities’ successes in the achievement of their students when tested at the age of seven – a process which is envisaged to develop into the publication of achievements at several ages and school by school).

Activity 2: Up to 45 minutes

Split the class into two main sections, though there may be more than one sub-group in each section. One section is to examine the strengths and weaknesses of norm-referencing, and the other group is to examine the strengths and weaknesses of criterion-referencing. They will have 20 minutes to prepare this. After that spend the remainder of the task on a feedback session. Split the whiteboard into four sections:

Strengths of norm-referencing	Strengths of criterion-referencing
Weaknesses of norm-referencing	Weaknesses of criterion-referencing

On the whiteboard, enter the feedback comments that they make, and then ask them what they notice, e.g.:

- Are the weaknesses of norm-referencing ‘put right’ by the strengths of criterion-referencing?
- Are the weaknesses of criterion-referencing ‘put right’ by the strengths of norm-referencing?
- To what extent are norm-referencing and criterion-referencing mutually exclusive?
- Is norm-referencing or criterion-referencing most widely used in the UK, and why?

Activity 3: 30 minutes

This is more of a preparation for session *three*. In that session there will be a debate on the motion ‘This house believes that there should be state-wide examinations at ages 16 and 18 in the UK. Divide the group into two. One half will be people supporting the motion, the other half will be people speaking against the motion. The group speaking for the motion will need to nominate two people to start the debate by speaking in favour of the motion and one person to summarise at the end. The group speaking against the motion will need to nominate two people to start the debate against the motion and one person to speak against it. The sequence of the debate will be:

Speaker one to speak in favour of the motion
 Speaker two to speak in favour of the motion
 Speaker one to speak against the motion
 Speaker two to speak against the motion

Open the debate to everyone

Closing speech by a speaker in favour of the motion

Closing speech by a speaker against the motion

Voting.

So, the activity in this session (one) is to move into groups, to nominate the speakers, and to start to agree the points to be made.

Handouts

(a) PowerPoint slides

(b) Module schedule.

PRESENTATION 2

This presentation continues the introduction to a range of different types of assessments.

Before you go to Slide 2, introduce domain-referencing thus:

An outgrowth of criterion-referenced testing has been the rise of domain-referenced tests. Here considerable significance is accorded to the careful and detailed specification of the content or the domain that will be assessed. The domain is the particular field of area of the subject that is being tested, for example, light in science, two-part counterpoint in music, parts of speech in English language. The domain is set out very clearly and very fully, such that the full depth and breadth of the content is established. Test items are then selected from this very full field, with careful attention to sampling procedures so that representativeness of the wider field is ensured in the test items.

Slide 1

Slide 2: Domain-referencing

The stages in setting up a domain-referenced assessment are set out in the slide. A test or assessment must commence by identifying the domain that it is sampling, which, as a general rule, should be specific rather than general. For example, the domain of 'music' is very general, and it is unlikely that one could devise an assessment that would cover all the aspects of music; so one would probably have a highly selective and unrepresentative sampling of the domain of 'music'. On the other hand, if one breaks down music into several components or smaller domains, then the chances of fairly covering each smaller domain are increased. The issue is one of identifying an appropriately small domain. It is the difference between being able to say, for example, that a student was generally musical, based on highly selective evidence, and saying that she was able to perform music by a range of baroque composers; in the former the domain is unhelpfully general and in the latter the statement tells us something about the student's performing ability, knowledge of the baroque repertoire and its variety. There are conceptual and ideological disputes about what constitutes a domain, and what its component elements are; for example: what constitutes literacy, numeracy, musicality, a skill, and a competence.

Domain-referencing frequently cites the three domain taxonomies of Bloom and his associates as examples of domains: cognitive, affective and psychomotor, within which there are several levels (e.g. in the cognitive domain there is a putative progression from the lowest level of knowledge, through comprehension, application, analysis, synthesis to evaluation at the highest order), with allocation of marks to different orders of cognitive thinking.

A domain-referenced assessment will need to define and delimit its domain(s) clearly and specifically (e.g. content areas, knowledge areas, skill areas). It will then sample items within each domain that are intended to be a fair representation of the whole domain. There would be the possibility of building up a profile of performance across domains,

and reporting that profile. The issue here is that the more specific, and hence reliable and informative, the domain-referenced assessment are, the longer and more detailed will be the assessment. It is a trade-off.

How to score the results is addressed in the next slide.

Slide 3

With regard to the scoring, the student's achievements on that test are computed to yield a proportion of the maximum score possible, and this, in turn, is used as an index of the proportion of the overall domain that she has grasped. So, for example, if a domain has 1,000 items and the test has fifty items, and the student scores thirty marks from the possible fifty then it is inferred that she has grasped 60 per cent $((30 \div 50) \times 100)$ of the domain of 1,000 items. Here inferences are being made from a limited number of items to the student's achievements in the whole domain; this requires careful and representative sampling procedures for test items.

It can be seen that for a domain to be specified usually required a long list of items to be drawn up. This means that domain construction and sampling are usually the stuff of experts at national levels, e.g. national examination bodies.

Slide 4

Domain-referencing is not without its problems, not the least of which is to secure agreement on what validly constitutes the domain itself. It will need to specify overall levels of difficulty, so that all the questions could be ranked in order of difficulty, or in terms of a logical sequence, such that one would not be able to answer later questions if one did not have the knowledge that was tested in earlier questions. However, attempts to identify such progression have proved elusive. Difficulty is a matter for subjective judgement, relative to individual learners, and they demonstrate their understanding sometimes inconsistently, though attempts to attenuate this problem can be made through trialling. Put simply, it is not easily possible to match clear criteria to grades except in a general, rather than specific, way. Grade-related criterion-referencing, either *a priori* or *post hoc*, have proved highly problematic.

The difficulties remain, then, with domain-referencing of: (a) identifying, agreeing and specifying the domains; (b) overcoming the problems of aggregation of scores from a range of questions; (c) overcoming the problems of identifying levels of difficulty; (d) agreeing the criteria and scoring of achievement of criteria for grading.

Slide 5: Diagnostic assessment

Diagnostic assessment is designed to identify particular strengths, weaknesses and problems in students' learning. Though it is often reserved for specialists (e.g. educational psychologists), this is by no means always the case, as teachers are constantly

diagnosing students' needs and problems. Diagnostic assessment is the foundation for formative assessment and planning, informing what a teacher should do next.

Diagnostic assessment is often used in the preparation of a statement for a student with special educational needs, describing exactly what the problems and needs are so that a programme of remediation can be focused and appropriate. Not that this applies only to children with difficulties; indeed, given the discussion in part two of teachers' apparent inabilities to match work to high achievers, there is a clear need for careful diagnosis of every student to take place in order that a better level of matching can be achieved. Diagnostic assessment is often an in-depth form of assessment.

Slide 6: Formative assessment

Formative assessment suggests and shapes the contents and processes of future plans for teaching and learning. Formative assessment – assessment *for* learning – provides feedback to teachers and students on students' current performances, achievements, strengths and weaknesses in such a form that it is clear what the student or the teacher can do next either to improve, enhance or extend learning and achievement. Assessment is to educate and improve learning and performance, not simply to audit it. In this sense formative assessment is constructive and useful; it sets an agenda for improvement. It takes place *during* a programme so that it can shape the forthcoming areas of the programme. Formative assessment can be frequent and informal, thereby really assisting teachers and students in the day-to-day business of improvement. It is designed to figure highly in planning for learning.

Slide 7

Effective teaching is flexible and adaptive, and requires the constant matching and adjustment of teaching programmes to meet learner's needs, not least in terms of pace and style. If assessment does not have any real impact on teaching and learning plans, e.g. if the teacher presses on regardless of the results of the feedback, then formative assessment has not really taken place. Indeed in the UK, the school inspectors (OFSTED) indicated the quality and use of ongoing assessment was good or better in only 37 per cent of schools inspected. Clearly there is a need for development here.

Formative assessment is often linked to teachers' own formal and informal assessments of students; it moves further than correcting mistakes to identifying causes of problems in learning. Whilst it can provide rich, detailed and day-to-day ongoing information about a student's performance, they may be prone to bias in the halo effect (where a teacher, for whatever reason, has a positive view of one aspect of a child and this affects all her views of that child) and in the teacher's own confusion – perhaps with the best of motives – between effort and achievement, rewarding high effort and industry even though the achievement may be poor. This might be acceptable in an assessment of effort, diligence or industry, but it has little place in an assessment of achievement.

Paul Black – a key assessment figure in the United Kingdom – makes the very telling point that one cannot have genuine or extended formative assessment unless one is prepared to modify the curriculum. This is a powerful message for those committed to a lock-step curriculum, whose pace, timing, and contents are prescribed for every student. He makes the point that formative assessment cannot just be bolted onto an existing scheme; it changes schemes. This has significant messages for the level of detail of pre-planning that can take place in classrooms, indeed it might confound the best laid plans for target setting by governments. One way of addressing this is to provide reinforcement activities for slower learners, and extension activities for faster learners. In short, formative assessment connotes, indeed requires, differentiated learning.

Slides 8 and 9

Black suggests that there are problems with teachers conducting their own formative *tests*, not least of which is that they resort to simplistic testing rather than richer and more extended forms of assessment. Indeed he cites four main problems:

- ‘Classroom evaluation practices generally encourage superficial and rote learning, concentrating on recall of isolated details, usually items of knowledge which pupils soon forget.
- Teachers do not generally review the assessment questions that they use and do not discuss them critically with peers, so there is little reflection on what is being assessed.
- The grading function is over-emphasised and the learning function is under-emphasised.
- There is a tendency to use a normative rather than a criterion approach, which emphasises competition between pupils rather than personal improvement of each. The evidence is that with such practices the effect of feedback is to teach the weaker pupils that they lack ability, so that they are de-motivated and lose confidence in their own capacity to learn.’

He also cites examples of situations where teachers collect vast quantities of information but are then unclear on what to do with it all – the purposes and criteria for the selection and use of assessment data have not been made clear. Indeed he suggests that teachers in fact may use such data to provide them with positive feedback on their own good performance rather than to assess students. Formative feedback and feedforward is much richer than simply administering tests and examinations.

Slide 10

Black makes the important point about setting and marking homework in this context, suggesting that if teachers calculate an average mark of 7 out of 10 then they will take this as an indication to continue rather than as an indication of the need to revise and re-teach, despite the fact that, by implication (an average implies scores above and below the average) a significant proportion of the class may not have understood half of what was taught.

Before Slide 11

There are several conditions under which formative assessment works well. These are linked to improvement, motivation and other affective factors (e.g. self-esteem), and reactive (i.e. the teacher adjusts her/his teaching as a consequence of the assessment), and which provide rich feedback and feedforward to students.

Formative assessments are integral to the everyday learning that takes place in schools. Formative assessments require the exercise of *judgement* about exactly what a student needs to do next and why, thereby respecting the professional judgement of teachers. Formative assessment and the feedback gained from it by students and teachers can also be a motivator, as it outlines successes and shows how weaknesses or failures can be remedied – it is designed to be positive, supportive and helpful. It recognises a student's positive achievements and builds on these. It is optimistic in indicating that there is a way forward.

Including teacher assessment in high stakes assessment is problematic. On the one hand teachers should be involved as they should have both the professional expertise and should be afforded the status that involvement in high stakes assessment brings. On the other hand ensuring reliability and validity in teacher assessments, and selecting from formative assessments, which is the stuff of teacher assessment, is problematic.

Slide 11

Black is unequivocal on teacher assessment: it is effective in raising levels of achievement and motivation if:

- it is criterion-referenced rather than norm-referenced;
- it uses praise rather than blame;
- it is differentiated to meet individual needs;
- it concentrates on, and is referenced to, learning goals;
- sets attainable targets;
- is part of a flexible and changeable programme of learning;
- students enter the 'frame of reference of the teacher'.

Further, he indicates that formative assessment is at its best when it involves the student on ipsative assessment and this is addressed later. In a major study Black and Wiliam report research (some 250 studies) that demonstrates that, despite methodological research difficulties, formative assessment does make a positive difference in yielding substantial learning gains. That said, they indicate that it involves considerable changes in classroom practices and paradigms of teaching in order to realise its maximum benefit.

Slide 12

Formative assessment should lead to rich, formative feedback to students, i.e. feedback on which they can know how to act to improve their learning and achievements, something which a mark or a grade simply does not have the power to do. Formative assessment with formative feedback to learners offers very great opportunities for them to improve, not least because it is made clear to them what they have to do to improve – the ‘rules of the game’ are made explicit. This was shown to advantage particularly those low-achieving students. Formative feedback on performance improves students’ motivation, their involvement, self-esteem and teacher-student relationships, all of which are integral to successful learning. Put simply, rich feedback improves performance. This is in sharp contrast to the awarding of marks and grades, which can be counterproductive in improving learning.

Formative feedback is feedback that relates intention to actuality – how far a student achieved his or her intentions, and what the gap was between what was desired and what was required, and the reasons for this gap, and what the gap was between ideal and actual performance. The best feedback is very specific, comments on what was actually done, is clear to the student, relates to targets, goals, and standards, and indicates specifically what has to be done to improve.

Ineffective, i.e. unhelpful feedback comprises statements like: ‘try harder’; ‘your spelling is poor’; ‘70 per cent’; ‘Grade D’. It often is given and received some time after the event, and it is too late to change the event itself (akin to gathering intelligence after the war is over). It is ‘one-off’ and restricted in its scope. It is impossible for the learner to know how to improve and what to do to improve. Effective feedback, on the other hand, provides such guidance, indicating what needs to be done to improve, what are the targets and how they can be reached, where attention needs to be focused, how errors can be corrected, how the learner can improve, and it is given in time for such improvements to be made, i.e. it is timely, frequent and ongoing. Feedback tells the student what were the results of her/his work; added to that, guidance – feedforward – enables the student to act on the feedback.

Slides 13 and 14

Formative assessment plays a major role in student learning. Improving learning through assessment is dependent on several key factors:

- ‘the provision of effective feedback to pupils;
- the active involvement of pupils in their own learning;
- adjusting teaching to take account of the results of assessment;
- a recognition of the profound influence assessment has on the motivation and self-esteem of pupils, both of which are crucial for learning;
- the need for pupils to be able to assess themselves and understand how to improve;
- sharing learning goals with pupils;
- involving pupils in self-assessment;
- providing feedback which leads to pupils recognising their next steps and how to take them;

- underpinned by confidence that every student can improve.’

Formative assessment, it is clear, is a major means of improving learning. For formative assessment to work it is important to share learning goals and intentions with the students, not least so that they know what is to be expected, and to separate clearly in the students’ minds what they have to learn (learning intentions and success criteria) from what they have to do to learn (the activities and tasks themselves). Inviting the students themselves to identify the success criteria is a useful mechanism to developing their own involvement in learning and assessment. Indeed Black and Wiliam argue that, for formative assessment to be productive, pupils should be ‘trained in self-assessment so that they can understand the main purposes of their learning and thereby grasp what they need to do to achieve’.

At issue here is the importance of sharing learning intentions with the students. This is critical. By such sharing it is suggested that the learners become more focused, motivated, engaged, persevering, self-motivated, well-behaved, owning of the learning, self-evaluative, and more enthusiastic about learning. This applies from the youngest children upwards; indeed 5-year-olds are able to suggest success criteria very successfully. What is evident here is that the success criteria to be used for evaluating performance and learning have to be made explicit for the learners, related to learning outcomes.

Further, formative feedback underlines the central importance of providing rich feedback to students. This moves away from simply orthographic feedback (e.g. on presentation, style, spelling, punctuation, quantity and handwriting) to qualitative feedback on how to improve (being, more strictly, ‘*feedforward*’ rather than simply *feedback*). Simply undertaking marking is often directly responsible for regression in learning, being demoralising for many learners. Indeed many learners, when given *both* a mark/grade and richer feedback, tend to ignore the feedback and only concentrate on the grade or mark, thereby neglecting the opportunity to improve. Similarly, giving too much summative feedback can be overwhelming. The trick is to provide ongoing, rich feedback to learners.

Formative feedback, within the context of clearly stated learning intentions, enables students to identify the gap between what they have actually done and what they could have done. It enables students to learn what they need to do to improve. It is premised on communication and dialogue between the teacher and the learner, and is designed to promote reflective thinking by learners. It is designed to improve motivation that, in turn, promoted effective learning. Formative assessment has a major contribution to make in the fields of: raising levels of motivation to learn; deciding what to learn; learning how to learn, and evaluating learning.

As such, it can be seen that formative assessment is closely linked to principles of metacognition.

Slide 15: Summative assessment

Formative assessment contrasts with *summative* assessment both in timing and purpose. Summative assessment – assessment *of* learning – is terminal; it comes at the end of a programme and assesses, for example, students’ achievements in the programme and of overall knowledge acquisition and practice. It is the stuff of formal examinations, the end of term test, the A level, the final examinations for a degree programme. A summative assessment provides data on what the student has achieved at the end of a course; it might also be more of a retrospective review of what has taken place during the course and what has been learned from it. Summative assessment is often concerned with certification, the awarding of marks and grades and public recognition of achievement.

Summative assessment is frequently used for the purpose of transition and transfer – to pass onto the next party in education (e.g. secondary school, sixth form college, university) information about achievement and progress, or indeed it can be used for *selection* (e.g. for higher education, for occupational routes, for vocational routes, for academic routes). Transfer between different stages of schooling, if communication is to be effective, requires a shared understanding between the givers and receivers of the information. This might necessitate standardised assessments (e.g. national tests). Such information must be: (a) *detailed* (e.g. a profile rather than simply a grade); (b) based on common *criteria* for grading (i.e. be consistent); (c) based on the same procedure for determining *standards* of grading (i.e. so that a grade B means the same for all parties). Quite how far such summative assessment will, in fact, be useful is a moot point, as new institutions often adopt different teaching and learning styles, different learning environments, different contents. To address this it might be useful to include assessment of ‘core’ or transferable’ skills, i.e. elements which cut across subject boundaries, and perhaps personality characteristics.

Slide 16

Summative assessment is often construed as testing, the strengths and weaknesses of which are examined later in the module. Summative assessment carries the major risk of a negative backwash effect on the curriculum, narrowing the curriculum to that which will appear on the assessment (often the examination) and narrowing the learning to a limited range of activities. Put simply, summative assessment can become behaviourist rather than embodying the more open-ended, constructivist view of learning.

Formative and summative assessment appear to lie in tension with each other, even though there may be some degree of overlap: certification and accountability in summative assessment steer assessment towards the production of simple grades which have little formative function – they are literally largely useless for planning teaching and learning, whilst formative assessment for planning day-to-day teaching and learning requires a much fuller, detailed kind of assessment, with a different purpose and focus.

Break

Activity 1: How to improve the extent and use of formative assessment

Time allowed: 45 minutes

Place the class into five or six smaller groups for discussion and feedback. The questions to be addressed in their discussions are:

- 1 How can the use of formative assessment in schools be increased?
- 2 What has to be done to improve formative assessment in schools?

Each group is to prepare answers to these two questions, on overhead transparency, and to be prepared to present four bullet points (only) as answers for each question, though the transparency may contain more for handout later. The feedback session gives three minutes to each group.

At the end of the presentations give out the summary sheets on the LEARN project (on involving pupils in assessment and on giving feedback) from the UK in 2000.

Activity 2:

The remainder of the evening will be spent in preparing the debate for Session 3.

Time allowed: 45 minutes

Handouts

- (a) PowerPoint slides
- (b) Summary of part of the LEARN project.

PRESENTATION 3: Ipsative assessment

Slide 1

Slide 2: Ipsative assessment

Ipsative assessment (derived from the Latin 'ipse' – meaning 'herself' or 'himself') has more than one interpretation. In the sense of *ipsative-referencing* it can refer to the principle that an individual's performance in a particular field of learning must be judged in relation to his or her other performances either in the same domain at different times or in different domains (to give an overall profile).

In another sense ipsative assessment refers to a process of self-assessment in which students identify their own starting points or, in the language of action planning and school development planning, *initial conditions*. This is undertaken *in the student's own terms* (hence the appeal to the Latin root). From this analysis the student sets targets for future learning and achievements, often in conjunction with the setting of a time frame. When the student arrives at the end of the nominated time frame she reviews her progress within that time and her levels of achievement of the targets that she set for herself. This accords with the view that students should be given the opportunity to reflect on their own learning. Ipsative assessment starts and finishes with the student setting the agenda; it may not relate to formal assessment in form, content, process or timing, but it is valuable in that the student has a degree of ownership of the assessment process. It is much less threatening to many students than external and formal assessments. Furthermore, it can focus on aspects of development that do not necessarily feature in many formal assessments, for example social, moral and emotional development.

Slide 3: Self-assessment

Self-assessment can involve, for example:

- 'reflecting on past experience;
- seeking to remember and understand what took place;
- attempting to gain a clearer idea of what has been learned or achieved;
- sharing responsibility for the organisation of their [students'] work;
- keeping records of activities they [students] have undertaken;
- making decisions about future actions and targets.'

Black, a celebrated researcher into assessment in the UK, is quite clear that students are perfectly capable of monitoring themselves, provided that they have been educated to do this. Indeed he draws a loose link between self-referencing, motivation, self-esteem and the development of metacognition. Here, it seems, is a comparatively straightforward device, which can improve learning on a variety of fronts. Self-assessment works particularly well when students are very clear on the purpose and focus of learning and on the criteria for judging successful learning. Training in self-monitoring, and its outcomes lead to successful learning and 'significant learning gains', not least because

learners become more reflective and independent (e.g. the development of metacognition). Further, there is considerable positive impact on students' self-esteem when they are involved in their own self-evaluation. Self-assessment promotes the management of students' own learning.

Ipsative assessment feeds directly into action planning, i.e. to plan how to improve, together with target setting. As with action planning that uses feedback from assessment formatively, ipsative assessment need not be undertaken by the student in isolation, rather it can be undertaken with a teacher as facilitator and negotiator.

Authentic assessment

Introduction

There has been a growing trend to render assessment more like real life, using real evidence from real situations, rather than through the use of proxy or surrogate indicators of achievement like tests. Students should be able to demonstrate their learning and abilities in real rather than contrived or narrow situations. To rely on written assessments alone is ridiculous: it is like asking a golfer to demonstrate her golfing ability by writing an essay about it; rather we should ask her to demonstrate her golfing abilities by actually playing golf! Similarly, it would be ridiculous to assess a person's driving ability only by a written test and without requiring him to actually drive, or to assess a person's musicianship without requiring her to actually play a note of music. Authentic assessment relates assessment to the real world of what people actually do rather than using some easy-to-score responses to questions. What makes authentic assessment authentic is that all parties know what students can actually do in real life with the knowledge, skills and competencies that they have learnt.

Black gives the example of the child being asked to handle real measuring equipment in science, rather than simply a drawing of it. Her results improve when using the actual equipment. Lest we are accused of seriously misrepresenting the actual nature of assessment in schools, let us be mindful of the vast sums of money spent each year administering solely written assessments, and the significance accorded to written assessments in judging students, not only in school but for the future, e.g. in university and employment entrance.

Many traditional assessments are frequently simplistic inasmuch as they require simple right answers, are frequently disconnected from the everyday world, are conducted through a print medium, contain fragmented and often relatively random knowledge, are conducted on a 'one-off' basis, yield a simple score, often do not declare the marking criteria and lead to naïve notions of passing and failing. This stands in stark contrast to the real world which is complex, contains actual use of knowledge in live contexts, integrates knowledge and competencies from a range of sources, iterative and developmental, requires rich and ongoing feedback for further learning, which concerns actual practice *in situ*, and which cannot be summarised in a singular score.

Slides 4, 5 and 6: Authentic assessment

There are several principles and characteristics of authentic assessment:

- Students are assessed on what has been taught and practised.
- An educative assessment and system is designed to improve learning and achievement and is based on meaningful, credible and realistic (real-world) tasks and activities.
- The focus of teaching and instruction is on solving problems and accomplishing tasks that are like those addressed by professionals in the field.
- Standards of performance are transparent – to the public and learners – and are the focus of feedback.
- Feedback is timely and ongoing.
- Tasks are authentic and provide performer-friendly feedback.
- It enables the teacher to observe processes as well as outcomes.
- It enables complex knowledge to be demonstrated in integrated tasks.
- It provides challenging yet realistic tasks.
- A hand-on exercise or problem is expected to be solved that produces an observable outcome or product.
- Assessment is linked closely to learning.
- Student engagement and motivation are addressed.
- The tasks cover fully the domains to be addressed.
- Tasks, learning and performance are on real projects, promoting motivation and engagement in real world projects.
- Assessment can last over a flexible period of time, during which time the teacher helps the student to improve.
- Learning contexts are provided in which students show what they have learned in the same conditions as those that occur in real life.
- Assessment is part of ongoing learning; it is built-in to learning, not a ‘bolt-on’ addendum.

‘Authentic instruction and assessment identify the knowledge, thinking, problem-solving skills, social skills, and attitudes expected by those in the community, on the job, or in advanced courses as part of their normal work. Authenticity involves testing what was taught and practised in class and asks learners to use the same skills, knowledge, and thought processes modelled by adults at work, presented in class activities, covered in the text and workbook, and required outside the classroom. Instruction and assessment are authentic when the tasks that are the focus of instruction and assessment are important tasks that tell the learners something about their knowledge and skills relative to themselves, rather than others. In other words, the learner’s grades represent what he can do on some important task, not how far above or below average he is. . . . Authentic instruction and assessment, therefore, are designed to produce the learner’s best, rather than typical, performance.’

Authentic assessment is concerned with the processes as well as the products of learning, with criterion-referencing, an increased role for self-assessment, and with mastery learning and application. Clearly the determination of tasks for assessment is critical here, to enable learning and application to be demonstrated and to ensure coverage of a range of domains. Indeed the boundaries of the domains and the assessment have to be delimited and communicated clearly. The nature of the evidence has to be clarified and communicated in authentic assessment (e.g. observation, demonstration, performance, outcome). Transparency is critical.

In conducting the assessment allowance has to be given for teacher intervention, e.g. in reducing the task requirements for students in difficulty, or for extending them for students who can handle it. Further, as many examples of authentic assessment activities involve collaborative work, clarification must be ensured on how each individual's contribution will be recognised (and how much the more assertive members of the group have given everyone the opportunity to demonstrate their learning). In many cases this might be difficult to disentangle, and it raises several problems, for example if one student's performance was a consequence of another's then how is this factored into the assessment?

What we have in authentic assessment is a major move towards increasing the validity of assessments, though the reliability is difficult to address. Even if criteria, marking schemes, grades and contents are made explicit, the problems come in applying these consistently, fairly and with equity and generalisability across different projects, students, teachers and contexts.

We commented in a previous session that the *purposes* of assessment not only are distinct but lie in tension with each – the more assessment serves one purpose the less it can serve another. This is true for some *types* of assessment. For example, the more we move towards summative, grade-related examinations the more we move away from formative and diagnostic assessments that required detailed – often qualitative – comments; the more we move towards standardised tests the more we move away from ipsative assessments; the more we move towards norm-referenced assessments that often yield a single score or grade the more we move away from criterion-referenced assessments that will yield specific details about a range of elements; the more we use external, objective instruments the less opportunity we have to use internal, teacher-devised instruments (often simply as a function of the time available). We suggest that student teachers should be concerned with diagnostic and formative assessments, providing useful feedback to students, i.e. feedback to improve learning, and that these can be part of ipsative assessment and action planning. Further, by involving students in ipsative assessment and by providing feedback to students upon which they can act to improve, positive interpersonal relationships are developed between students and student teachers that, themselves, support enhanced learning through engagement and motivation.

Activity 1: 45 minutes

In small groups, to address two questions:

- 1 Does anything need to be done in the UK to make assessment more authentic?
- 2 If so, what are the problems to be overcome in making assessment more authentic in the UK?

The preparation of the responses should take no more than 20 minutes, and the feedback should give each group around 3 minutes each to present their summaries (onto the whiteboard) and their results.

Break

The session after the break is one single activity. It is to conduct a debate, based on the motion 'This house believes that there should be state-wide examinations at ages 16 and 18 in the UK', which was signalled in Session 1.

The motion is 'This house believes that there should be state-wide examinations at ages 16 and 18 in the UK'. Write it on the whiteboard. Remind the group that the sequence of the debate is:

Speaker one to speak in favour of the motion
 Speaker two to speak in favour of the motion
 Speaker one to speak against the motion
 Speaker two to speak against the motion
 Open the debate to everyone
 Closing speech by a speaker in favour of the motion
 Closing speech by a speaker against the motion
 Voting.

Conduct the debate.

Handout

(a) PowerPoint slides

PRESENTATION 4

Introduction

The next three presentations examine the issue of tests. The first two of these look at test construction, and the third looks at the strengths and weaknesses of testing in the UK. Testing is widespread all over the world, and some of its strengths and weaknesses are examined here, together with issues in test construction and design.

The major means of gathering assessment data over the years have been tests and examinations. Published tests are commercially produced and they take various forms: *diagnostic* tests (e.g. the Metropolitan Diagnostic Tests), *aptitude* tests (which predict a person's aptitude in a named area, e.g. the Comprehensive Test of Adaptive Behaviour, the McCarthy Screening Test, the Assessment for Training and Employment Test), *achievement* tests, *norm-referenced* tests (the Boehm Test of Basic Concepts), *criterion-referenced* tests (e.g. the GCSE examinations of course work), *reading* tests (e.g. the Edinburgh Reading Test), *verbal reasoning* tests (e.g. the Wechsler Adult Intelligence Scale and tests published by the National Foundation for Educational Research), tests of *critical thinking* (e.g. the Watson-Glaser Critical Thinking Appraisal), tests of *social adjustment* (e.g. the British Social Adjustment Test and the Kohn Social Competence Scale), *baseline assessment* tests (e.g. the Basic Achievement Skills Individual Screener). Several commercial companies hold tests that have restricted release or availability, requiring the teacher or school to register with a particular company. For example, in the United Kingdom the Psychological Corporation Ltd not only holds the rights to a worldwide battery of tests but also has different levels of clearance for different users. Having different levels of clearance attempts to ensure, for example, that students are not 'prepared' for the test by coaching on the various items.

Published tests have several attractions: they are objective and standardised (as a result of piloting and refinement), they declare their levels of reliability and validity through the inclusion of statistical data, they come complete with instructions for administration and processing, they are often straightforward and quick to administer and mark, and an accompanying manual gives guidance for the interpretation of results.

On the other hand, simply *because* they have been standardised on a wide population and are generalisable, by definition they are not tailored to an individual institution, a local context or specific needs. Hence if published tests are to be used they must serve the desired purposes of the assessment – the notion of *fitness for purpose* is crucial in selecting from a battery of tests in the public domain.

Slide 1

Slide 2: Components of test construction

A test that is devised by the teacher for a specific purpose, whilst it does not have the level of standardisation of a commercially produced test, nevertheless is tailored to that

teacher's particular needs, addressing very fully the notion of *fitness for purpose*. Guidelines for the construction of a test by a teacher require attention to be given to:

- the *purposes* of the test must be explicit (i.e. to provide data for a particular type of assessment);
- the *type* of test must be appropriate (e.g. diagnostic, achievement, aptitude, criterion-referenced, norm-referenced)
- the *objectives* of the test need to be stated in operational terms;
- the *content* of the test must be suitable;
- the *construction* of the test must address *item analysis* (e.g. ensuring that each item in the test serves one or more specified objectives), *item discriminability* and *item difficulty* (see the discussion of these topics below);
- the *format, readability and layout* of the test must be appropriate and clear for students, its layout, instructions, required methods of working and of completion;
- the *validity and reliability* of the test must be appropriate (see the discussion above of these two areas);
- the *marking criteria and marking conventions* for the test must be explicit, including the weightings of various elements;
- the *instructions* for the test, including the administration, marking and data treatment of the test.

We unpack these in this session and the next.

Slide 3

Teacher-devised tests are prone to several problems:

- they encourage rote/superficial learning, simply for the test day itself;
- discussions of reliability, validity and utility are not undertaken between teachers – the tests are a private creation by individual teachers;
- quantity can be favoured over quality;
- there is a tendency to lead to an over-emphasis on marks and grades, to the detriment of learning and rich feedback;
- they foster too much of a competitive mentality in learners;
- they lead to 'learned helplessness' in students (as discussed in module two) – where they are only motivated by the desire to do well rather than to learn and where they come to believe that if they fail then it is because they are not sufficiently clever and so there is nothing that can be done about it to improve, therefore they avoid risk-taking and challenge;
- too much testing can be counterproductive, leading to a decline in performance, particularly if the results of the test are simply a mark or grade rather than rich feedback on how to improve;
- grading every piece of work is simply counterproductive.

Slide 4

In devising a test the teacher will have to consider several stages:

1 Identify the purposes of the test

The purposes of a test are several, for example to *diagnose* a student's strengths weaknesses and difficulties, to measure *achievement*, to measure *attainment*, to measure *aptitude* and *potential*, to identify *readiness* for a programme ('placement testing') and is usually a form of pre-test, for *formative*, *diagnostic* or *summative* purposes.

2 Identify the test specifications

The test specifications include:

- which programme objectives and student learning outcomes will be addressed;
- which content areas will be addressed;
- the relative weightings, balance and coverage of items;
- the total number of items in the test;
- the number of questions required to address a particular element of a programme or learning outcomes;
- the exact items in the test.

To ensure validity in a test it is essential to ensure that the objectives of the test are fairly addressed in the test items. Objectives should: (a) be specific and be expressed with an appropriate degree of precision; (b) represent intended learning outcomes; (c) identify the actual and observable behaviour which will demonstrate achievement; (d) include an active verb; (e) be unitary (focusing on one item per objective).

One way of ensuring that the objectives are fairly addressed in test items can be done through a matrix frame that indicates the *coverage* of content areas, the coverage of *objectives* of the programme, and the *relative weighting* of the items on the test. Such a matrix is set out below, taking the example from a secondary school history syllabus.

Give out these two examples and comment on them as below:

Weighting a matrix of test items

<i>Content areas</i>	<i>Objective/area of programme content</i>			<i>Objective/area of programme content</i>			<i>Objective/area of programme content</i>			<i>Total</i>
	1a	1b	1c	2a	2b	2c	3a	3b	3c	
Aspects of the 1939–45 war										
The build-up to the 1939–45 world war	1	2		2	1	1	1	1	1	10
The invasion of Poland	2	1	1	3	2	2	3	3	3	20
The invasion of France	3	4	5	4	4	3	4	4	4	35
The allied invasion	3	2	3	3	4	3	3	2	2	25
The end of the conflict	2	1		1	1	1	2	2		10
<i>Total</i>	11	10	9	13	12	10	13	12	10	100

This indicates the main areas of the programme to be covered in the test (*content areas*); then it indicates which objectives or detailed content areas will be covered (1a to 3c) – these numbers refer to the identified specifications in the syllabus; then it indicates the marks/percentages to be awarded for each area. This indicates several points:

- the least emphasis is given to the build-up to and end of the war (10 marks each in the ‘total’ column);
- the greatest emphasis is given to the invasion of France (35 marks in the ‘total’ column);
- there is fairly even coverage of the objectives specified (the figures in the ‘total’ row only vary from 9–13);
- greatest coverage is given to objectives 2a and 3a, and least coverage is given to objective 9c;
- some content areas are not covered in the test items (the blanks in the matrix).

Hence we have here a test scheme that indicates relative weightings, coverage of objectives and content, and the relation between these two latter elements. Relative weightings should be addressed by firstly assigning percentages at the foot of each column, then by assigning percentages at the end of each row, and then completing each cell of the matrix within these specifications. This ensures that appropriate sampling and coverage of the items are achieved. The example of the matrix refers to specific objectives as column headings; of course these could be replaced by factual knowledge, conceptual knowledge and principles, and skills for each of the column headings. Alternatively they could be replaced with specific aspects of an activity, for example: designing a crane, making the crane, testing the crane, evaluating the results, and improving the design. Indeed these latter could become content (row) headings thus (below):

Weighting elements of test items

<i>Content area</i>	<i>Identifying key concepts and principles</i>	<i>Practical skills</i>	<i>Evaluative skills</i>	<i>Recording results</i>	<i>Total</i>
Designing a crane	2	1	1	3	7
Making the crane	2	5	2	3	12
Testing the crane	3	3	1	4	11
Evaluating the results	3		5	4	12
Improving the design	2	2	3	1	8
<i>Total</i>	12	11	12	15	50

Here one can see that practical skills will carry fewer marks than recording skills (the column totals), and that making and evaluating carry equal marks (the row totals).

This exercise also enables some indication to be gained on the number of items to be included in the test, for instance in the example of the history test above the matrix is $9 \times 6 = 54$ possible items, and in the 'crane' activity above the matrix is $5 \times 4 = 20$ possible items. Of course, there could be considerable variation in this, for example more test items could be inserted if it were deemed desirable to test one cell of the matrix with more than one item (possible for cross-checking), or indeed there could be fewer items if it were possible to have a single test item that serves more than one cell of the matrix. The difficulty in matrix construction is that it can easily become a runaway activity, generating very many test items and, hence, leading to an unworkably long test – typically the greater the degree of specificity required, the greater the number of test items there will be. One skill in test construction is to be able to have a single test item that provides valid and reliable data for more than a single factor.

Having undertaken the test specifications, the researcher should have achieved clarity on (a) the exact test items that test certain aspects of achievement of objectives, programmes, contents etc.; (b) the coverage and balance of coverage of the test items; and (c) the relative weightings of the test items.

3 Select the contents of the test

Make it clear that you will be returning to this matter in just a moment, but you want to present a complete list of matters in summary form before you go into details. Hence the next slides continue and complete the list then you return to issues of the contents of the test. This enables all the items of test construction to be kept together.

Slide 5

4 Consider the form of the test

Many tests are of the pen-and-paper variety. Clearly this need not be the case, for example tests can be written, oral, practical, interactive, computer-based, dramatic, diagrammatic, pictorial, photographic, involve the use of audio and video material, presentational and role-play, and simulations. The form of the test will need to consider, for example, reliability and validity, difficulty, discriminability, marking and grading, item analysis, timing. Indeed several of these factors take on an added significance in non-written forms of testing; for example: (a) reliability is a major issue in judging live musical performance or the performance of a gymnastics routine – where a ‘one-off’ event is likely; (b) reliability and validity are significant issues in group performance or group exercises – where group dynamics may prevent a testee’s true abilities from being demonstrated. The teacher will need to consider whether the test will be undertaken individually, or in a group, and what form it will take. The test will need to consider the presentation (introduction) mode, the activity mode (what will be done in the test) and the response (outcome) mode – what product the student is expected to produce.

5 Write the test item

We consider these later.

6 Consider the layout of the test

This will include:

- (i) The nature, length and clarity of the instructions (e.g. what to do, how long to take, how much to do, how many items to attempt, what kind of response is required (e.g. a single word, a sentence, a paragraph, a formula, a number, a statement etc.), how and where to enter the response, where to show the ‘working out’ of a problem, where to start new answers, e.g. in a separate booklet), is one answer only required to a multiple choice item, or is more than one answer required.
- (ii) Spread out the instructions through the test, avoiding overloading students with too much information at first, and providing instructions for each section as they come to it.
- (iii) What marks are to be awarded for which parts of the test?
- (iv) Minimising ambiguity and taking care over the readability of the items.
- (v) The progression from the easy to the more difficult items of the test (i.e. the location and sequence of items).
- (vi) The visual layout of the page, for example, avoiding overloading students with visual material or words.
- (vii) The grouping of items – keeping together items that have the same contents or the same format.

(viii) The setting out of the answer sheets.

The layout of the text should be such that it supports the completion of the test and that this is done as efficiently and as effectively as possible for the student.

7 Consider the timing of the test

This refers to two areas: (a) when the test will take place (the day of the week, month, time of day, and (b) the time allowances to be given to the test and its component items. With regard to the former, in part this is a matter of reliability, for the time of day, week etc. might influence how alert, motivated, capable a student might be. With regard to the latter, the researcher will need to decide what time restrictions are being imposed and why (for example, is the pressure of a time constraint desirable – to show what a student can do under time pressure – or an unnecessary impediment, putting a time boundary around something that need not be bounded – was Van Gogh put under a time pressure to produce the painting of sunflowers?).

Though it is vital that the student knows what the overall time allowance is for the test, clearly it might be helpful to a student to indicate notional time allowances for different elements of the test; if these are aligned to the relative weightings of the test (see the discussions of weighting and scoring) they enable a student to decide where to place emphasis in the test – she may want to concentrate her time on the high scoring elements of the test.

8 Plan the scoring of the test

The awarding of scores for different items of the test is a clear indication of the relative significance of each item – the weightings of each item are addressed in their scoring. It is important to ensure that easier parts of the test attract fewer marks than more difficult parts of it; otherwise a student's results might be artificially inflated by answering many easy questions and fewer more difficult questions. Clearly, also, it is important to know in advance what constitutes a 'good answer'.

The more marks that are available to indicate different levels of achievement (e.g. for the awarding of grades), the greater the reliability of the grades will be, though, clearly this could make the test longer. Scoring will also need to be prepared to handle issues of poor spelling, grammar and punctuation – is it to be penalised, and how will consistency be assured here? Further, how will issues of omission be treated, e.g. if a student omits the units of measurement (miles per hour, dollars or pounds, metres or centimetres)?

Related to the scoring of the test is the issue of reporting the results. If the scoring of a test is specific then this enables variety in reporting to be addressed, for example, results may be reported item by item, section by section, or whole test by whole test. This degree of flexibility might be useful for the student teacher, as it will enable particular strengths and weaknesses in groups of students to be exposed.

Underpinning the discussion of scoring is the need to make it unequivocally clear exactly what the marking criteria are – what will and will not score points. This requires a clarification of whether there is a ‘checklist’ of features that must be present in a student’s answer. The specification of the performance criteria is crucial, defining high scoring, medium scoring and low scoring criteria. In essence, what is being required here is a rubric for the test, specifying:

- The performance criteria.
- What to look for in judging performance.
- The range of quality of performance and how to score different levels of performance.
- How to determine reliability and validity, and how these are reflected in the scoring;

Clearly criterion-referenced tests will have to declare their lowest boundary – a cut-off point – below which the student has been deemed to fail to meet the criteria. A compromise can be seen in those criterion-referenced tests which award different grades for different levels of performance of the same task, necessitating the clarification of different grade cut-off points in the examination.

9 Consider special adaptations to the test.

There will be some students who will need to have special arrangements made for them for the test, for example in terms of:

- the presentation format of the test (the need to read them slowly to a student, or in stages rather than all at the start, or to have them read aloud, or language levels adjusted, or, indeed, not to have them in written form);
- the response format (allowing dictionaries and calculators, allowing responses other than in written form, having a scribe to write the dictated answers, allowing the use of notes);
- the timing of the test (providing extra time, avoiding timed tests, providing breaks through different parts of the test, allowing an unlimited amount of time);
- the setting of the test (testing in a separate place, providing a one-to-one test situation, reducing distractions).

Slide 6: Stages of a test

In considering the stages of the test it important to consider:

- How will the task be introduced (e.g. oral, written, pictorial, computer, demonstration).
- What will the students be doing when they are working (e.g. mental work, practice, oral work, written work, making something).
- What will the outcome be (e.g. multiple choice answer, short response, essay, oral, written, practical, computer output, artifact, table of results).

Slide 7: Operationalising a test

In moving to test construction the teacher will need to consider how each element to be tested will be *operationalised*: (a) what indicators and kinds of evidence of achievement of the objective will be required; (b) what indicators of high, moderate and low achievement there will be; (c) what will the students be doing when they are working on each element of the test; (d) what the outcome of the test will be (e.g. a written response, a tick in a box of multiple choice items, an essay, a diagram, a computation). Attention will have to be given to the *presentation, operation* and *response* modes of a test: (a) how the task will be introduced (e.g. oral, written, pictorial, computer, practical demonstration); (b) what the students will be doing when they are working on the test (e.g. mental computation, practical work, oral work, written); and (c) what the outcome will be – how they will show achievement and present the outcomes (e.g. choosing one item from a multiple choice question, writing a short response, open-ended writing, oral, practical outcome, computer output). Operationalising a test from objectives can proceed by stages:

- identify the objectives/outcomes/elements to be covered;
- break down the objectives/outcomes/elements into constituent components or elements;
- select the components that will feature in the test, such that, if possible, they will represent the larger field (i.e. domain referencing, if required);
- recast the components in terms of specific, practical, observable behaviours, activities and practices that fairly represent and cover that component;
- specify the kinds of data required to provide information on the achievement of the criteria;
- specify the success criteria (performance indicators) in practical terms, working out marks and grades to be awarded and how weightings will be addressed;
- write each item of the test;
- conduct a pilot to refine the language/readability and presentation of the items, to gauge item discriminability, item difficulty and distracters (discussed below), and to address validity and reliability.

Slides 8 and 9: Item analysis

An item analysis will need to consider:

- the suitability of the format of each item for the (learning) objective (appropriateness);
- the ability of each item to enable students to demonstrate their performance of the (learning) objective (relevance);
- the clarity of the task for each item;
- the straightforwardness of the task;
- the unambiguity of the outcome of each item, and agreement on what that outcome should be;
- the cultural fairness of each item;

- the independence of each item (i.e. where the influence of other items of the test is minimal and where successful completion of one item is not dependent on successful completion of another);
- the adequacy of coverage of each (learning) objective by the items of the test;
- are all the items in the test equally difficult;
- which items are easy, moderately hard, hard, very hard;
- what kinds of task each item is addressing (e.g. is it (a) a practice item – repeating known knowledge, (b) an application item (applying known knowledge, (c) a synthesis item – bringing together and integrating diverse areas of knowledge);
- if not, what makes some items more difficult than the rest;
- whether the items are sufficiently within the experience of the students;
- how motivated students will be by the contents of each item (i.e. how relevant they perceive the item to be, how interesting it is).

Item analysis is designed to ensure that: (a) the items function as they are intended, for example, that criterion-referenced items fairly cover the fields and criteria and that norm-referenced items demonstrate *item discriminability* (discussed later); (b) the level of difficulty of the items is appropriate (discussed later); (c) the test is reliable (free of distracters – unnecessary information and irrelevant cues). An item analysis will consider the accuracy levels available in the answer, the item difficulty, the importance of the knowledge or skill being tested, the match of the item to the programme, and the number of items to be included.

Slide 10: Item discriminability and item difficulty

In constructing a test the researcher will need to undertake an item analysis to clarify the item discriminability. In other words, how effective is the test item in showing up differences between a group of students? Does the item enable us to discriminate between students' abilities in a given field? An item with high discriminability will enable the researcher to see a potentially wide variety of scores on that item; an item with low discriminability will show scores on that item poorly differentiated. Clearly a high measure of discriminability is desirable.

Distracters are the stuff of multiple choice items, where incorrect alternatives are offered, and students have to select the correct alternatives. Here a simple frequency count of the number of times a particular alternative is selected will provide information on the effectiveness of the distracter: if it is selected many times then it is working effectively; if it is seldom or never selected then it is not working effectively and it should be replaced.

Slide 11: Item difficulty

If we wished to calculate the *item difficulty* of a test, we could use the following formula:

$$\frac{A}{N} \times 100$$

where

A = the number of students who answered the item correctly;

N = the *total* number of students who attempted the item.

Hence if twelve students out of a class of twenty answered the item correctly, then the formula would work out thus:

$$\frac{12}{20} \times 100 = 60\%$$

The maximum index of difficulty is 100 per cent. Items falling below 33 per cent and above 67 per cent are likely to be too easy and too difficult respectively. It would appear, then, that this item would be appropriate to use in a test. Here, again, whether the student teacher uses an item with an index of difficulty below or above the cut-off points is a matter of judgement. In a norm-referenced test the item difficulty should be around 50 per cent.

Ethical issues in preparing for tests

A major source of unreliability of test data derives from the extent and ways in which students have been prepared for the test. These can be located on a continuum from direct and specific preparation, through indirect and general preparation, to no preparation at all. With the growing demand for test data (e.g. for selection, for certification, for grading, for employment, for tracking, for entry to higher education, for accountability, for judging schools and teachers) there is a perhaps understandable pressure to prepare students for tests. This is the 'high-stakes' aspect of testing where much hinges on the test results. At one level this can be seen in the backwash effect of examinations on curricula and syllabuses; at another level it can lead to the direct preparation of students for specific examinations. Preparation can take many forms:

- ensuring coverage, amongst other programme contents and objectives, of the objectives and programme that will be tested;
- restricting the coverage of the programme content and objectives to those only that will be tested;
- preparing students with 'exam technique';
- practice with past/similar papers;
- directly matching the teaching to specific test items, where each piece of teaching and contents is the same as each test item;
- practice on an exactly parallel form of the test;

- telling students in advance what will appear on the test;
- practice on, and preparation of, the identical test itself (e.g. giving out test papers in advance) without teacher input;
- practice on, and preparation of, the identical test itself (e.g. giving out the test papers in advance), with the teacher working through the items, maybe providing sample answers.

How ethical it would be to undertake the final four of these is perhaps questionable, or indeed any apart from the first on the list. Are they cheating or legitimate test preparation? Should one teach to a test; is not to do so a dereliction of duty (e.g. in criterion- and domain-referenced tests) or giving students an unfair advantage and thus reducing the reliability of the test as a true and fair measure of ability or achievement? In high stakes assessment (e.g. for public accountability and to compare schools and teachers) there is even the issue of not entering for tests students whose performance will be low. There is a risk of a correlation between the 'stakes' and the degree of unethical practice – the greater the stakes, the greater the incidence of unethical practice. Unethical practice occurs where scores are inflated but reliable inference on performance or achievement is not, and where different groups of students are prepared differentially for tests, i.e. giving some students an unfair advantage over others. To overcome such problems, she suggests, it is ethical and legitimate for teachers to teach to a broader domain than the test, that teachers should not teach directly to the test, and the situation should only be that better instruction rather than test preparation is acceptable.

One can add to this list of considerations the view that:

- (a) tests must be valid and reliable;
- (b) the administration, marking and use of the test should only be undertaken by suitably competent/qualified people (i.e. people and projects should be vetted);
- (c) access to test materials should be controlled, for instance: test items should not be reproduced apart from selections in professional publication; the tests should only be released to suitably qualified professionals in connection with specific professionally acceptable projects;
- (d) tests should benefit the testee;
- (e) clear marking and grading protocols should exist (the issue of transparency);
- (f) test results are only reported in a way that cannot be misinterpreted;
- (g) the privacy and dignity of individuals should be respected (e.g. confidentiality, anonymity, non-traceability);
- (h) individuals should not be harmed by the test or its results.

Break

The rest of the evening will be to introduce and to start planning the assessment activity.

Handouts

- (a) PowerPoint slides
- (b) Weighting elements of a test

(c) The assignment activity

PRESENTATION 5

Introduction

This session introduces some key aspects of constructing items for a test.

Many tests are composed of a number of items – for example, missing words, incomplete sentences or incomplete, unlabelled diagrams, true/false statements, open-ended questions where students are given guidelines for how much to write (e.g. a sentence, a paragraph, 300 words etc.), closed questions, multiple choice questions, matching pairs of statements and responses, true-false items, short answer and long answer responses. They can test recall, knowledge, comprehension, application, analysis, synthesis, and evaluation, i.e. different orders of thinking. These take their rationale from the work of Bloom in 1956 on hierarchies of cognitive abilities – from low order thinking (comprehension, application) to higher order thinking (evaluation), which were introduced in module one.

Slide 1

Slide 2

This slide presents a list of kinds of items that might be included in a test. These will be unpacked in what follows, with reference to handout material.

Then go through the handout material, giving examples where relevant.

Missing words and incomplete sentences

Missing words items are useful for rapid completion, and guessing is reduced because a specific response is required. On the other hand such items tend to require lower level recall and can be time-consuming to score, if the marker has to try to understand what the student has been thinking about in writing the answer.

The test will need to address the intended and unintended clues and cues that might be provided in it, for example:

- the number of blanks might indicate the number of words required;
- the number of dots might indicate the number of letters required;
- the length of blanks might indicate the length of response required;
- the space left for completion will give cues about how much to write;
- blanks in different parts of a sentence will be assisted by the reader having read the other parts of the sentence (anaphoric and cataphoric reading cues);

There are several guidelines for constructing short-answer items to overcome some of these problems:

- make the blanks close to the end of the sentence;
- keep the blanks the same length;
- require a single word or short statement for the answer;
- ensure that there can be only a single correct answer;
- omit only the key words and avoid omitting so many key words as to make the sentence unintelligible, maybe making the blanks appear towards the end of the sentence;
- avoid putting several blanks close to each other (in a sentence or paragraph) such that the overall meaning is obscured;
- only make blanks of key words or concepts, rather than of trivial words;
- avoid addressing only trivial matters;
- ensure that students know exactly the kind and specificity of the answer required;
- specify the units in which a numerical answer is to be given;
- use short-answers for testing knowledge recall.

Multiple choice statements

Multiple choice items can test lower order and higher order thinking. They are quick to complete and to mark; they are objective and are widely used in formal tests, though they may take some time to devise. In devising fixed, closed response questions there are several considerations to be borne in mind, for example:

- make the question and requirements unambiguous and in a language appropriate for the students;
- avoid negatives in statements;
- avoid giving clues in the 'wrong' choices to which is the correct response;
- provide around four choices in order to reduce guessing, and ensure that the 'distracters' are sufficiently close to the correct response as to be worthy of consideration by the student, i.e. make the options realistic;
- keep the choices around the same length;
- try to avoid negative statements;
- avoid giving grammatical cues in the choices (e.g. the word 'an' in the stem requires an option that begins with a vowel; the word 'is' in the stem requires an option written in the singular);
- ensure that one option does not contain more information than another, as this suggests to students that this is the correct option;
- avoid the use of 'all of the above' or 'none of the above', as these tend to become the options chosen;
- avoid value and opinion statements, as these are contestable;
- consider the use of pictures, tables, graphics and maps, particularly for higher order multiple choice questions (e.g. the siting of a supermarket).

There are several attractions to multiple choice items, for example:

- they can be completed quite rapidly, enabling many questions to be asked which, in turn, enable good coverage of each domain, thereby increasing reliability and validity;
- there is limited writing, so students' writing skills (or their lack) do not impede demonstration of knowledge;
- the opportunities for errors or biases in marking are reduced;
- many papers can be scored at speed (e.g. through optical mark scanners);
- they yield numbers (e.g. of correct responses), which can be converted into statistics to inform accountability issues.

On the other hand they have attracted severe criticism:

- they demean and reduce the complexity of knowledge, learning and education to the trivial, atomised and low level;
- they have little diagnostic or formative potential;
- scores may be inflated through informed guessing;
- there is no indication that correct choices are made for the correct reason;
- they lead to teaching to the test, particularly in high stakes situations;
- they only address issues to which there is a putative right answer.

In devising tests attention has to be given to the issue of student choice: what is deemed to be unavoidably and centrally important, over which there is no option, and what might be optional is a matter for decision. Black suggests that offering students choices of questions in a test does not help them to achieve higher marks, whilst it engages issues of reliability (e.g. consistency of demand across questions) and validity (e.g. to cover the required domains to be tested). Indeed some students may make unwise choices, and this might compromise reliability. The issue of choice extends further, to include whether one gives different tests to different students, for example, depending on their anticipated performance, or whether one offers a single graduated test, with items becoming progressively more difficult the further one moves through the test. Black alludes to a potential gender issue here, in that girls may not choose some difficult items in mathematics as they do not want to take such risks, even though in fact they may have the ability to undertake them.

With regard to multiple choice items there are several potential problems:

- the number of choices in a single multiple choice item (and whether there is one or more right answer(s));
- the number and realism of the distractors in a multiple choice item (e.g. there might be many distractors but many of them are too obvious to be chosen – there may be several redundant items);
- the sequence of items and their effects on each other;
- the location of the correct response(s) in a multiple choice item.

There are several suggestions for constructing effective multiple choice test items:

- ensure that they catch significant knowledge and learning rather than low-level recall of facts;
- frame the nature of the issue in the stem of the item, ensuring that the stem is meaningful in itself (e.g. replace the general ‘sheep’: (a) are graminivorous, (b) are cloven footed, (c) usually give birth to one or two lambs at a time’ with ‘how many lambs are normally born to a sheep at one time?’);
- ensure that the stem includes as much of the item as possible, with no irrelevancies;
- avoid negative stems to the item;
- keep the readability levels low;
- ensure clarity and unambiguity;
- ensure that all the options are plausible so that guessing of the only possible option is avoided;
- avoid the possibility of students making the correct choice through incorrect reasoning;
- include some novelty to the item if it is being used to measure understanding;
- ensure that there can only be a single correct option (if a single answer is required) and that it is unambiguously the right response;
- avoid syntactical and grammatical clues by making all options syntactically and grammatically parallel and by avoiding matching the phrasing of a stem with similar phrasing in the response;
- avoid including in the stem clues as to which may be the correct response;
- ensure that the length each response item is the same (e.g. to avoid one long correct answer from standing out);
- keep each option separate, avoiding options which are included in each other;
- ensure that the correct option is positioned differently for each item (e.g. so that it is not always option 2);
- avoid using options like ‘all of the above’ or ‘none of the above’;
- avoid answers from one item being used to cue answers to another item – keep items separate.

True-false items

True-false items are useful in being quick to devise and complete and are easy to score. They offer students a fifty-fifty chance of being correct simply by guessing. To overcome the ‘guess factor’ students could be asked to indicate why a false item is false, indeed to re-write it to make it true, where possible, with marks being awarded for the correct revision. In constructing true-false items it is important to ensure that the statements are unequivocal, unambiguously true or false. This means omitting questions of value or questions in which there may be differences of opinion. It is important to keep the statements of approximately the same length, to avoid extremes (e.g. ‘never’, ‘only’, ‘always’) as these will not be chosen, and to avoid double negatives.

There are particular problems in true-false questions:

- ambiguity of meaning;

- some items might be partly true or partly false;
- items that polarise – being too easy or too hard;
- most items might be true or false under certain conditions;
- it may not be clear to the student whether facts or opinions are being sought;
- as this is dichotomous, students have an even chance of guessing the correct answer;
- an imbalance of true to false statements;
- some items might contain ‘absolutes’ which give powerful clues, e.g. ‘always’, ‘never’, ‘all’, ‘none’.

To overcome these problems there are several points that can be addressed:

- avoid generalised statements (as they are usually false);
- avoid trivial questions;
- avoid negatives and double negatives in statements;
- avoid over-long and over-complex statements;
- ensure that items are rooted in facts;
- ensure that statements can be either only true or false;
- write statements in everyday language;
- decide where it is appropriate to use ‘degrees’ – ‘generally’, ‘usually’, ‘often’ – as these are capable of interpretation;
- avoid ambiguities;
- ensure that each statement only contains one idea;
- if an opinion is to be sought then ensure that it is attributable to a named source;
- ensure that true statements and false statements are equal in length and number.

Matching items

The use of matching items is another rapid-to-construct and easy-to-score means of testing, and are useful for measuring associations between statements and facts. They keep the role of guessing to a minimum. Matching items comprise a descriptions list and an options list, with the options list to be matched to the appropriate descriptions list. In writing lists of matching items it is important to:

- keep each of the two lists of items homogeneous;
- ensure that the options are plausible distracters;
- ensure that the descriptions list contains longer phrases or statements than the options list;
- provide clear instructions for how to indicate the matching (e.g. by joining lines, by writing a number and a letter);
- ensure that there are more options than descriptions, to address the issue of distracters;
- indicate in the instructions whether the options can be used more than once.

There are also particular potential difficulties in matching items:

- it might be very clear to a student which items in a list simply *cannot* be matched to items in the other list (e.g. by dint of content, grammar, concepts), thereby enabling the student to complete the matching by elimination rather than understanding;
- one item in one list might be able to be matched to several items in the other;
- the lists might contain unequal numbers of items, thereby introducing distracters – rendering the selection as much a multiple choice item as a matching exercise.

Difficulties in matching items can be addressed thus:

- ensure that the items for matching are homogeneous – similar – over the whole test (to render guessing more difficult);
- avoid constructing matching items to answers that can be worked out by elimination (e.g. by ensuring that: (a) there are different numbers of items in each column so that there are more options to be matched than there are items; (b) students can avoid being able to reduce the field of options as they increase the number of items that they have matched; (c) the same option may be used more than once);
- decide whether to mix the two columns of matched items (i.e. ensure, if desired, that each column includes both items and options);
- sequence the options for matching so that they are logical and easy to follow (e.g. by number, by chronology);
- avoid over-long columns and keep the columns on a single page;
- make the statements in the options columns as brief as possible;
- avoid ambiguity by ensuring that there is a clearly suitable option that stands out from its rivals;
- make it clear what the nature of the relationship should be between the item and the option (on what terms they relate to each other);
- number the items and letter the options.

Essay questions

A more open-ended type of written assessment is an essay. It is the freedom of response that is possible in the essay form of examination that is held to be its most important asset, enabling higher order and sustained, in-depth and complex thinking to be demonstrated.

With regard to essay questions, there are several advantages that can be claimed. For example, an essay, as an open form of testing, enables complex learning outcomes to be measured, it enables the student to integrate, apply and synthesise knowledge, to demonstrate the ability for expression and self-expression, and to demonstrate higher order and divergent cognitive processes. Further, it is comparatively easy to construct an essay title. On the other hand, essays have been criticised for yielding unreliable data, for being prone to unreliable (inconsistent and variable) scoring, neglectful of intended learning outcomes and prone to marker bias and preference (being too intuitive, subjective, holistic, and time-consuming to mark).

Slides 3 and 4

To overcome these difficulties the authors suggest that:

- instructions must be given as to whether the requirement is for a short or long essay;
- the essay question must be restricted to those learning outcomes that are unable to be measured more objectively;
- the essay question must ensure that it is clearly linked to desired learning outcomes; that it is clear what behaviours the students must demonstrate;
- the essay question must indicate the field and tasks very clearly (e.g. ‘compare’, ‘justify’, ‘critique’, ‘summarise’, ‘classify’, ‘analyse’, ‘clarify’, ‘examine’, ‘apply’, ‘evaluate’, ‘synthesise’, ‘contrast’, ‘explain’, ‘illustrate’);
- time limits are set for each essay;
- options are to be avoided, or, if options are to be given, ensure that, if students have a list of titles from which to choose, each title is equally difficult and equally capable of enabling the student to demonstrate achievement, understanding etc.;
- marking criteria are prepared and are explicit, indicating what must be included in the answers, what cognitive processes are being looked for in the essay (e.g. higher order and lower order thinking) together with their specification in the essay requirement (e.g. ‘compare’, ‘speculate’, ‘contrast’, ‘evaluate’, ‘give reasons for’) and the points to be awarded for such inclusions or ratings to be scored for the extent to which certain criteria have been met;
- decisions are agreed on how to address and score irrelevancies, inaccuracies, poor grammar and spelling;
- the scoring criteria are agreed, e.g. for content, organisation, logic, structure, presentation, secretarial skills, reasonableness, coverage, completeness, internal consistency, originality, creativity, level of detail, persuasiveness of the argument, conclusiveness, clarity, demonstration of understanding and application, and so on;
- the marks to be awarded for each element are agreed, including the weighting of the marks;
- the work is marked blind, and, where appropriate, without the teacher knowing the name of the essay writer. Of course this is perhaps difficult or even undesirable for the student teacher, who may need to know the identity of the writer. The issue here is how to avoid personal knowledge clouding an objective judgement.

There are disadvantages, however, in the essay as a gatherer of information. Essays are difficult to assess reliably. With only one or two assessors a considerable degree of unreliability can creep into the assessment of essays, i.e. ‘inter-rater’ reliability may be limited. Even with one marker there is a high degree of unreliability. Since only a limited number of essay titles can be answered in one examination, the candidate can address only a limited part of a syllabus of work. The student who has the misfortune to choose the ‘wrong’ essay title may produce work that does not fairly represent her true abilities.

There are some ways of overcoming these weaknesses in the essay form of test. First, all students might be asked to write on the same essay title(s), the principle being that individuals can only be compared to the extent that they have 'jumped the same hurdles'. Second, marking should be *analytic* rather than *impressionistic*. Analytic marking is based upon *prior decisions* about what exactly is being assessed in the essay – the content? The style? The grammar? The punctuation? The handwriting? (i.e. criterion-referencing should apply). On the question of the low agreement between essay markers, there are several ways of reducing the subjective element in marking:

- by marking for substantive content rather than style;
- by fixing a maximum penalty for mistakes in grammar, syntax and spelling;
- by multiple marking followed by a further close scrutiny of those essays placed near the pass-fail line.

Break

The remainder of the evening is to continue the planning of the assessment activity. At the end of the evening make the point that the assessment is largely in the form of a test containing several items, but that there may be other forms of assessment to be included, and some of these will feature in the next session.

Handouts

- (a) PowerPoint slides
- (b) Sheets on writing test items.

PRESENTATION 6

Introduction

Make the point that this is quite an intensive evening in terms of having a longer than usual initial presentation. The intention is to introduce some wider forms of assessment that feed into the assessment activity that is being planned.

Slide 1

Slides 2 and 3: Performance assessment

A related aspect of authentic assessment is performance assessment, indeed there is some interchange of terminology here, as they share several common characteristics.

Performance assessment:

- concerns direct reality rather than disconnected items of knowledge;
- models the real learning that students undertake rather than contrives artificial tests; it does not distort teaching;
- requires students to demonstrate what they can do rather than simply completing test items which address fragments of what they might do and reporting what they could or would do in a particular situation, i.e. it is actual rather than speculative;
- integrates many areas of knowledge in undertaking and demonstrating a particular activity or project;
- uses real-world activities and learning;
- uses activities which relate to the world outside school (e.g. vocational experience);
- is focused on processes as well as products and outcomes;
- can replace contrived test situations with everyday, ongoing teaching and learning activities and tasks, which are used for assessment purposes (though specific tasks can be set specifically for assessment purposes, i.e. whilst these might be excluded from authentic assessment they may not be excluded from performance assessment).

Performance assessment, as its name suggests, is that assessment which is undertaken of activities or tasks in which students can demonstrate their learning through performance in real situations. It typically uses teacher assessment, typically observation, questioning and professional judgement, rather than objective assessment and often uses some form of portfolio assessment (discussed later).

Performance assessment is already widespread in some subjects, for example, communication skills, psychomotor skills (e.g. physical education and athletics, music making, drawing, science experiments, design and technology, project work, drama, social skills in group activities). For example, instead of simply describing a science experiment, the students should actually do it with real equipment and real tasks. A performance assessment, as mentioned above, can be a deliberately structured task or test for assessment purposes, so long as it uses 'real world' activities and learning rather than, for example, a proxy assessment by a pen and paper test. It requires the learner to

demonstrate knowledge, learning and understanding through a real task and application. Hence performance assessment should strive to be as close to authentic assessment as possible. The choice of task is critical here, to provide students with the opportunity to demonstrate and apply their learning, often within severe classroom constraints such as time and resources.

Slide 4

Performance assessment will also need to take account of how open or closed the task is, i.e. how much freedom can be given to students to design and undertake tasks, to select resources and ways of working or, by contrast, whether they are highly constrained here. Similarly, account will have to be taken of the nature and amount of teacher intervention in the activity. Clearly the degrees of freedom and autonomy here will need to be reflected in the nature of the marking and scoring, maybe by awarding more marks for greater initiative and autonomy.

The introduction, activity and response (outcome) modes of the task will need to be considered here, and these are discussed below. For example, will the task be introduced orally, in writing, individually, in groups, to the whole class, in a standard format or in a variety of modes? What kind of activity will be required, and what kinds of thinking and behaviour will be required, for example: higher order thinking, group activity, paired work, individual project work, a practical task, a written task? What kind of outcome is required, for example: a written report, a project, an essay, a drawing, a plan, an oral presentation, a led discussion, a multi-media presentation, solutions to a problem, an artefact, numerical data and results, summary statements, a portfolio?

In planning for performance assessment it is important to plan for the instructions to be given, the marking criteria scheme to be used and communicated to the students (see below on test construction and scoring) and how evidence for assessment will be collected, e.g. formally in writing, or informally through aural and visual observation and questioning, or a combination of methods. Further, the nature of the feedback and feedforward to be given to the students is important, as is the effect of teacher intervention in the process. There are several questions to be addressed in considering performance assessment, for example:

- Is the marking going to involve grading, written feedback, oral feedback, summative assessment, formative assessment, ongoing or on a limited number of occasions, or what?
- How is a group project going to be assessed?
- How is account to be given for differential effort and input into a group project?
- How is account to be taken of individual abilities in an individual's task?

Slide 5

- How much does the output/outcome reflect the processes that went into the production of the outcome?

- How should assessment of the processes be undertaken?
- How are reliability and consistency going to be addressed across different activities, outcomes and different students?
- What are the performance criteria?
- What is the performance evidence and how is it going to be gathered?
- What is going to be assessed: products, processes, knowledge, understanding, application of knowledge, initiative, creativity, lower order thinking, higher order thinking, problem-solving, attitudes to learning, skills, competencies, reasoning?

Though there is a range of questions to be addressed in considering performance assessment, this should not detract from the real benefits to be obtained from this style of assessment; it is rich in authenticity, real-world value, and it provides opportunities for students to be assessed on what they have learned and what they can do, ‘for real’.

Slides 6 and 7

In planning a performance assessment there are several stages that can be followed:

- Step 1: Decide on a specific subject area.
- Step 2: Define cognitive processes and social skills you want to assess.
- Step 3: Design a task and task context (including consideration of goal relevance for the learners, levels of difficulty, multiple goals, multiple solutions, self-determined learning, and the clarity of directions).
- Step 4: Specify the scoring rubrics (measuring the goals, selecting an appropriate scoring system, assigning point values).
- Step 5: Identify important implementation considerations: identifying testing constraints of time, reference material, other people’s input, equipment, scoring criteria; delivering the assessment (introducing and structuring the task, motivation, coaching, independent work, debriefing).

A critical aspect of performance assessment is to decide in advance what constitutes the evidence for making judgements of performance, and how the evidence will be acquired, e.g. through writing, questioning, observation, presentation.

Though there is much to commend performance assessment, not least that it is strong on authenticity and validity, one should be mindful that it is time-consuming, complex to score and difficult to standardise (hence of questionable reliability) and of limited generalisability unless a wide range of tasks is sampled. In this respect it may be easier to use for formative and diagnostic purposes – assessment for learning – rather than summative purposes – assessment of learning. If performance assessment is to be used for comparing students and for high stakes assessment then this will require: (a) clear specification of (cognitive demands, scoring criteria, levels of performance and learning contexts; (b) clear calibration of scoring; (c) moderation procedures, together with their associated training of markers and audit of marks.

Performance assessment is also linked to portfolio assessment, and we move to this now.

Portfolios and samples of work

Slide 8

Authentic assessment may draw on portfolio assessment. A portfolio is a collection of pieces of students' work, which indicate accomplishments over time and context, which may be used to represent their best achievements, as they contain the samples of the best work and which best represent their development. Portfolios, compiled by the student, with or without the support from, and negotiation with, the teacher, are powerful ways of involving students, a form of ipsative assessment, on which we have commented above.

Slides 9 and 10

Portfolios are useful in that they:

- indicate best accomplishments;
- help students to evaluate themselves;
- indicate improvement and development over time;
- comprise ongoing assessment;
- integrate teaching and assessment;
- promote worthwhile, meaningful learning;
- provide a richer picture of a student's accomplishments than provided by tests, grades, report forms and single 'one-shot' assessments;
- connect to real life and real-world learning, tasks and activities;
- involve higher order thinking and reflection as well as lower order skills;
- provide feedback to interested parties about rates of progress, suitability of the curriculum and a student's development;
- place an emphasis on processes of learning as well as outcomes;
- encourage students to reflect on what constitutes good performance;
- can provide diagnostic information for students and teachers;
- encourage the all-round development of the student, including attitudes, likes, hopes, and feelings;
- develop student motivation and involvement;
- are targeted towards individual development and take account of individual needs and abilities.

The contents of a portfolio are not fixed, but they change over time and to suit different purposes and audiences, as work is selected in and selected out. Clearly it is important to know the purpose of the portfolio, whether, for example, it is intended to represent the *best* work of the student or the *typical* work of the student, as this affects the selection of the contents of the portfolio. Reflecting on the choice of samples for the portfolio is important, as it encourages a student to reflect on her/his best/poorest/average/easiest/hardest piece of work and the reasons for this judgement, together with considerations of the pieces of work that demonstrate the greatest improvement and progress.

Slide 11

A portfolio has many purposes and uses, for example:

- to act as a showcase of a student's best work, as selected by the student;
- to act as a showcase of a student's best work, as selected by the teacher;
- to reflect the student's interests;
- to chart development, improvement and rates of progress;
- to reflect the student's growing self-reflection and self-evaluation;
- to provide evidence of knowledge, skills, competencies, understandings, and application;
- to report on real-world learning and application;
- to act as a document of record.

Slides 12 and 13

In using portfolios for assessment it is important to decide:

- the purpose of the portfolio (e.g. to monitor progress, to communicate what has been learnt, to inform employers, to document achievements, to grade students, to select students for employment and higher education);
- the cognitive, affective and psychomotor skills, social skills, competencies, attitudes to be addressed in the portfolio;
- who will plan the portfolio (e.g. the teacher, the student, both, the parents);
- the contents of the portfolio and the sample of work to be chosen (e.g. the best work, typical work, a range of work, a close focus on a few items);
- the marking and scoring rubrics and processes, together with their communication and transparency;
- the weighting of the scores to be aggregated into a composite single mark or grade;
- the administrative procedures for the portfolios: the time frames for the sampling and deadlines for submission of items; procedures for the submission, grading and return of work; criteria for selection of work; storage of the portfolio; access to the contents of the portfolio;
- conferencing during, and as follow-up to, the preparation and final submission of the portfolio.

Slides 14 and 15

In marking/scoring portfolios several issues should be considered:

- select the performance to be taught;
- state the performance criteria: what constitutes effective learning, what are the learning outcomes, the required performance, the required activities and the required targets to be met;

- identify how students and teachers will be involved in the selection of the items for inclusion in the portfolio, the review of, and reflection on, the selection, and the opportunities to be provided for students to: (a) select in and out samples of work; (b) rework a particular activity or piece of work;
- state the number of performance levels (e.g. poor, average, above average);
- describe in detail the criteria for each of the performance levels;
- select the scoring level which most closely represents the student's performance and then award the grade;
- consider how best to communicate the feedback and marking criteria to students, and how to involve students in the assessment of the portfolio.

It is important for all parties to know the 'rules of the game' in devising and assessing portfolios, so that the criteria for inclusion and marking are transparent. There is the issue of whether it is acceptable to include poor samples of work, as students may have the right to privacy and to include only those samples of work which demonstrate their best achievements rather than the processes which led up to those achievements (should students be marked on drafts or poor work, or only on their best work?). The matter is akin to writing Curriculum Vitae: we deliberately exclude our failures and weaknesses, and only include those items that present ourselves in the best light.

Slide 16

Portfolios are not without their difficulties, for example:

- decisions on whether to include the best work or typical work (which may not be the best work);
- honesty (students may download materials from the internet, and pass them off as their own);
- time (portfolios take a long time to score and mark);
- storage and access (ensuring security and privacy);
- subjectivity (how to allow for subjectivity in the students and in the markers and how to address these without having everything double-marked);
- reliability (there is a need to score in sections and parts, i.e. analytically rather than, or as well as, holistically): how to allow for different choices in assignments and samples; how to allow for different formats of the portfolio.

Clearly these are tricky problems. At heart the portfolio is almost inevitably subjective and personal; the difficulty of reconciling this for use in a more objective style of assessment, for comparisons and comparative judgements of students is problematic.

Throughout the years of schooling samples of work can be used to provide assessment data, be they pieces of course work specifically undertaken for assessment purposes (for example, the course work elements for public examinations) or pieces of work undertaken as part of the everyday learning of students. Using samples of work for assessment represents one of the most widely used means of assessment because the samples of work are ongoing and rooted in the reality of everyday classroom life.

Ensuring that the samples of work fit the objectives and purposes of the assessment means that the criteria for setting the work in the first place must fit the assessment purposes, and that the assessment or marking of the work must make explicit the criteria to be used and disclose these to the students. In the interests of good teaching and natural justice there is little justification for withholding from students the purposes of the written work and the criteria that will be used to assess the work. It is unacceptable to have students 'play a game' whose rules they have not been told. Students will be more likely to feel involved in the process of assessment if it is made clear to them what it will be and what criteria will be used. This breaks with the traditional type of assessments where contents and criteria were kept secret.

Because portfolios are usually selections of best practice, and because it may not always be clear how much help from parents and teachers has gone into the portfolio or how much the student has simply downloaded material from the web, their reliability, indeed their validity, may be suspect, hence they tend not to be used for high stakes assessment. Indeed a decision has to be taken on whether to include *typical* or *best* work in the portfolio. Portfolios are very time-consuming to develop, review and mark, and this is a major problem for time-pressed teachers. Further, given that the contents of a portfolio may vary from one student to another, it becomes almost impossibly difficult to assess them fairly – consistently and comparatively – across students. Hence they tend to be used formatively rather than summatively. Portfolios are developmental and are intended to facilitate the process of communication and conferencing between teachers, learners, parents and maybe employers.

As an addition to the points presented here, give out the handout on marking work.

Break

After the break the remainder of the evening is devoted to the students continuing their preparation of the assessment activity for presentation in session eight.

Handouts

- (a) PowerPoint slides
- (b) Sheets of guidance on marking work.

PRESENTATION 7

Reliability and validity in assessments

Introduction

As educators we need reliable data on students' achievements so that we can have confidence both in how we judge students and in what we subsequently plan for students. Black reports that there was up to a 30 per cent chance of some students being placed in the wrong level in assessments. Reliability is compromised when students of the same ability and achievements score different results on the same test, when the same student scores differently on different tests of the same matters/contents, or when the same student scores differently on the same (or very similar) test on a different occasion. Reliability means that the results are consistent and reproducible with different markers, occasions, test items, test types, marking conventions, grading procedures and contexts.

Further, we need to be assured that the assessments that are made of students actually assess what they are intended to assess or else subsequent planning begins from the wrong place, i.e. the validity of assessment must be strong.

Slide 1

Slide 2

Reliability is an index of consistency and dependability, for example of marking practices/conventions and of standards. An assessment would have little reliability if it yielded different results in the hands of another assessor or different results for similar students. Reliability then requires comparability of practices to be addressed. This can be undertaken prior to assessments by agreement trials, so that a range of assessors can be clear and can agree on the specific marks and grades to be awarded for particular samples of work, examination scripts, course work and marks scored in elements of an overall assessment, though *in practice* it often only becomes an issue in the post-assessment standardisation and agreement of marks and awards. Reliability, then, affects the *degree of confidence* that one can put on assessment data and their interpretation.

Reliability is an important issue given the significant role of teachers in formal assessments and examinations, and the need for external markers of examination scripts to be fair to all candidates, neither too harsh nor too generous in comparison with other external examiners.

Not only must reliability be addressed but it must be seen to be addressed; marking must be seen to be fair and transparent. It comes as no surprise, therefore, that the significance of reliability and transparency should lead to objective, standardised, national, externally marked tests. Reliability in teachers' assessment can be improved by, amongst other things:

- joint planning between teachers in the same year or department, across years or across key stages;
- using the programme(s) of study to agree objectives for teaching, learning and assessment;
- developing common activities focused on agreed objectives;
- discussing and marking work to develop shared expectations of performance;
- comparing the performance of pupils from different classes on common activities;
- referring to national tests, tasks and published exemplification of standards;
- agreeing standards of samples of work from a range of contexts relating to particular level descriptions or end of key stage assessments;
- developing a common understanding of judgements about the work of individual pupils;
- identifying inconsistencies in pupils' performance;
- referring to examples of work whose standards have already been agreed and which are held in a school portfolio of work for facilitating moderation and consistency of grading;
- attending meetings outside the school for confirming judgements about standards, i.e. using external referents – for moderation purposes.

It can be seen merely from the size of this (incomplete) list that reliability features highly when assessment is undertaken externally and by teachers. It enters the assessment arena at the point of agreeing marks, i.e. after the product – the examination script or the course work for assessment, for example – has been made.

Activity 1: 45 minutes

Split the class into five or six groups to identify a range of threats to reliability. They should address possible threats in three main areas (write these on the whiteboard):

- threats to reliability in the markers/examiners themselves;
- threats to reliability from the students and teachers;
- threats to reliability from the assessment items;

Allow 20 minutes for the discussion and then have each group present back: three minutes maximum feedback for each group (in bullet points only).

Use this activity as a lead into the slides that follow.

Slide 3

There are several threats to reliability in assessments, for example: with respect to examiners and markers:

- errors in marking (e.g. attributing, adding and transfer of marks);

- inter-rater reliability (different markers giving different marks for the same or similar pieces of work);
- inconsistency in the examiner/marker (e.g. being harsh in the early stages of the marking and lenient in the later stages of the marking of many scripts);
- variations in the award of grades for work that is close to grade boundaries (some markers placing the score in a higher or lower category than other markers);
- the Halo effect, wherein a student who is judged to do well or badly in one assessment is given undeserved favourable or unfavourable assessment respectively in other areas.

Slide 4

With reference to the students and teachers themselves, there are several sources of unreliability:

- Motivation and interest in the task has a considerable effect on performance. Clearly, students need to be motivated if they are going to make a serious attempt at any test that they are required to undertake, where motivation is *intrinsic* (doing something for its own sake) or *extrinsic* (doing something for an external reason, e.g. obtaining a certificate or employment or entry into higher education). The results of a test completed in a desultory fashion by resentful pupils are hardly likely to supply the student teacher with reliable information about the students' capabilities. Research suggests that motivation to participate in test-taking sessions is strongest when students have been helped to see its purpose, and where the examiner maintains a warm, purposeful attitude toward them during the testing session.
- The relationship (positive to negative) between the assessor and the assessee exerts an influence on the assessment. There is sufficient research evidence to show that both *test-takers* and *test-givers* mutually influence one another during examinations, oral assessments and the like. During the test situation, students respond to such characteristics of the evaluator as the person's sex, age and personality. Although the student teacher can do little about his/her sex and age, it is important (and may indeed at times be comforting) to realise that these latent identities do exert potent influence. It could well be, for example, that the problems experienced by a female student conducting a test with older secondary school boys have little if anything to do with the quality of the test material or the amount of prior preparation she has put into the testing programme.
- The conditions – physical, emotional, social – exert an influence on the assessment, particularly if they are unfamiliar. The advice generally given in connection with the location of a test or examination is that the test room should be well lit, quiet and adequately ventilated. To this we would add that, wherever possible, students should take tests in familiar settings, preferably in their own form rooms under normal school conditions. Research suggests that distractions in the form of extraneous noise, walking about the room by the examiner, and intrusions into the room, all have significant impact upon the scores of the test-takers, particularly when they are younger pupils. An important factor in reducing students' anxiety and tension during an examination is the extent to which they are quite clear about what exactly they are

required to do. Simple instructions, clearly and calmly given by the examiner, can significantly lower the general level of tension in the test room. Student teachers who intend to conduct testing sessions may find it beneficial in this respect to rehearse the instructions they wish to give to pupils *before* the actual testing session. Ideally, test instructions should be simple, direct and as brief as possible.

- The Hawthorne effect, wherein, in this context, simply informing a student that this is an assessment situation will be enough to disturb her performance – for the better or the worse (either case not being a fair reflection of her usual abilities).
- Distractions (including superfluous information on the examination sheets).
- A considerable and growing body of research in the general area of *teacher expectancies* suggests that students respond to the *teacher-assessor* in terms of their perceptions of what he expects of them. It follows, then, that the calm, well-organised student teacher embarking purposefully upon some aspect of evaluation probably induces different attitudes (and *responses*) among her class of children than an anxious, ill-organised colleague.
- The time of the day, week, and month will exert an influence on performance. Some students are fresher in the morning and more capable of concentration.
- Students are not always clear on what they think is being asked in the question; they may know the right answer but not infer that this is what is required in the question.

Slide 5

- The students may vary from one question to another – a student may have performed better with a different set of questions that tested the same matters. Black argues that ‘two questions which seem to the expert to be asking the same thing in different ways might well be seen by the pupil as completely different questions’.
- Teachers teach to the test. This is perhaps unsurprising in high stakes assessment, or where, as in some countries, teachers’ contract renewal is contingent on students’ test results, or where ‘league tables’ of overall performance are published.
- Teachers and students practice test-like materials. There are entire lucrative businesses operating to prepare students for public tests, e.g. the GMAT and GRE tests in the USA, where entrance to university depends on test scores.
- A student may be able to perform a specific skill in a test but not be able to select or perform it in the wider context of learning.
- Cultural, ethnic and gender background affect how meaningful an assessment task or activity is to students, and meaningfulness affects their performance.
- Marking practices are not always reliable, teachers maybe being too generous, marking by effort and ability rather than performance.
- The context in which the task is presented affects performance: some students can perform the task in everyday life but not under test conditions.

Slide 6

With regard to the assessment items themselves, there may be problems (e.g. test bias), for example:

- The task itself may be multi-dimensional, for example, testing ‘reading’ may require several components and constructs. Students can execute a mathematics operation in the mathematics class but they cannot perform the same operation in, for example, a physics class; students will disregard English grammar in a science class but observe it in an English class. This raises the issue of the number of contexts in which the behaviour must be demonstrated before a criterion is deemed to have been achieved. The question of transferability of knowledge and skills is also raised in this connection. The *context* of the task affects the student’s performance.
- The validity of the items may be in question (discussed below).
- The language of the assessment and the assessor exerts an influence on the assessee, for example if the assessment is carried out in the assessee’s second language or in a ‘middle class’ code.
- The readability level of the task can exert an influence on the assessment, e.g. a difficulty in reading might distract from the purpose of an assessment, which is to test the use of a mathematical algorithm.
- The size and complexity of numbers or operations in an assessment (e.g. of mathematics) – that might distract the assessee who actually understands the operations and concepts.
- The number and type of operations and stages to a task – a student might know how to perform each element, but when they are presented in combination the size of the task can be overwhelming.
- The form and presentation of questions affects the results, giving variability on students’ performance.
- A single error early on in a complex sequence may confound the later stages of the sequence (within a question or across a set of questions), even though the student might have been able to perform the later stages of the sequence, thereby preventing the student from gaining credit for all she can, in fact, do.
- Questions requiring use of mechanical toys might favour boys more than girls.

Slide 7

- Questions requiring use of dolls or kitchen work might favour girls more than boys.
- Essay questions favour boys if they concern impersonal topics and girls if they concern personal and interpersonal topics.
- Boys perform better than girls on multiple choice questions and girls perform better than boys on essay-type questions (perhaps because boys are more willing than girls to guess in multiple choice items), and that girls performed better in written work than boys.
- Goulding indicated that continuous assessment of course work at 16-plus enabled a truer picture of girls’ achievements in mathematics to be presented than that yielded by results on a written examination. Girls may be more anxious about examinations than boys, and consequently their performance may suffer.
- Questions and assessment may be culture-bound: what is comprehensible in one culture may be incomprehensible in another.
- The test may be so long, in order to ensure coverage, that boredom and loss of concentration may impair reliability.

What we are saying is that specific contextual factors can exert a significant influence on learning and that this has to be recognised in conducting assessments, rendering an assessment as unthreatening and natural as possible.

Validity

Validity in assessment is defined as ensuring that the assessment in fact assesses what it purports to assess and provides a fair representation of the student's performance, achievement, potential, capabilities, knowledge skills etc. (i.e. addressing what it was intended to address). This is a problem when defining and operationalising abstract constructs like intelligence, creativity, imaginativeness, and anxiety. Validity refers to appropriateness, meaningfulness, usefulness, specificity, diagnostic potential, inferential utility and adequacy.

Slide 8

Face validity requires the assessment to assess what it was intended to assess.

Content validity requires the assessment to cover the intended contents in sufficient depth and breadth so as to be fair and adequate, and not to exceed the scope of those boundaries of content (i.e. not to cover items or contents that were not included in the programme).

Consequential validity depends on the way in which the results are used, i.e. that they should be used in the ways intended and not in other ways. Consequential validity would be violated if inferences made from the results of assessment were not sustainable or justified by the results themselves and were illegitimate. This requires the user of the results to know what the intentions of the assessment were. Of course, this is frequently violated when sensationalist headlines in the media indicate falling standards, when, in fact it was not possible to infer this legitimately from the data.

Predictive validity concerns how much the results of an assessment can be used to predict achievements in the future, e.g. how much scores at A level might be fair indicators of future degree classification. Low predictive validity (e.g. using A level scores to predict degree classification, where it is lower than 50 per cent) suggests that limited credence should be placed in such uses.

Construct validity requires the assessment to provide a fair operationalisation of the construct – often abstract – in question, e.g. intelligence, creativity, spatial awareness, problem-solving. This is usually the most difficult aspect of validity to address, not least because opinion is divided as to what a fair construction of the construct actually is. For example, exactly what intelligence is, and what proxy indicators of intelligence might be, can founder at the starting line if there is disagreement on whether it is a single ability, a multiple ability (e.g. Gardner's 'multiple intelligences'), a composite, innate or capable of being developed (nature or nurture).

One statistical means of addressing construct validity is to seek inter-correlations between several items that are intended to measure the same construct (or to undertake factor analysis, itself based on inter-correlations). The principle here is that inter-correlations between items in a test, for example, that are intended to measure the same construct should be higher than inter-correlations between items that are not intended to measure the same construct or which are intended to measure different constructs. Further, different types of question that are intended to measure the same construct should have stronger inter-correlations than inter-correlations using the same types of question to assess different constructs.

So here we have a dilemma. The more we steer towards reliability, consistency, uniformity, standardisation and grades, the more we move away from the rich data upon which teachers can often take action. Conversely, the more we move towards teacher- and student-defined, personalised valid data, the less generalisable, standardisable, comparable and consistent are the results (though no less transparent provided that the criteria are made public). Not only are reliability and validity sometimes in a state of tension with each other but that a third factor – manageability – might reduce reliability and validity. Manageability, reliability and validity may be in tension with each other.

Slide 9

Drawing together the several strands of the arguments and issues raised in this part we suggest several principles that should guide assessment.

Slide 10

These issues suggest several implications:

- the purposes are to be diagnostic and formative, providing feedback and being educative;
- teaching should be adjusted in light of assessment evidence;
- assessment should promote, not damage, student motivation and self-esteem;
- assessment should be constructively critical and provide rich, positive feedback and feedforward;
- the assessments should be criterion-referenced and the criteria should be public;
- the assessments should lead to diagnostic teaching;
- assessment should promote student self-evaluation.

Slide 11

- the assessments should be built on evidence rather than on intuition;
- assessment data should be derived from everyday classroom activities;
- assessment opportunities should be sought in everyday classroom activities;
- semi-structured approaches to gathering data are recommended, generating words rather than numbers (measures);

- assessments should be linked to the student teacher's and the student's action planning and target setting;

Slide 12

- involve the students in the assessment process;
- communicate the assessment criteria to students;
- demonstrate validity and reliability in the assessments, addressing particularly 'consequential validity';
- demonstrate *fitness for purpose* in deciding the method(s) of gathering assessment data and setting assessment tasks;
- select assessment methods that accord strongly with everyday teaching and learning processes.

Break

For the remainder of the session this is the last opportunity for students to prepare their presentations and associated handouts.

Handout

(a) PowerPoint slides