ISBN 1-903815-10-X

ltsn
Learning and Teaching
Support Network

LTSN Physical Sciences Practice Guide

Effective Practice in Objective Assessment

Alex Johnstone

Effective Practice in Objective Assessment

The Skills of Fixed Response Testing

Alex Johnstone

Published by the LTSN Physical Sciences Centre

Introduction

Assessment forms an integral part of the teaching and learning process. It is applied during a course (formative assessment) to help students and teachers to take a realistic view of progress and to catch misunderstandings at a stage when they can be easily rectified. It is also applied at the end of a course (summative assessment) to ensure that the student has not only learned the bits, but can also piece the bits together into a coherent whole.

This guide is largely concerned with the form of assessment called Objective or Fixed- Response testing, but this is only part of the range of assessment tools available. It would be a mistake to think that this form of testing is useful or appropriate for the measurement of all the skills we hope to develop in our students. However, objective testing has clear advantages for the assessment of some skills, not least its ease of scoring.

To clarify where these advantages lie, we shall have to consider the assessment process more widely in the early part of the guide.

The bulk of the guide is devoted to the development of the expertise necessary to design questions, to construct papers and to handle the results. The effort expended to gain this expertise will pay off in terms of better and sharper assessment tools which will help our students to take a realistic view of their progress and which will keep us from self-delusion about our teaching.

The examples of questions in the text have been purposely kept at a low content level to illustrate the method under discussion. This should not be taken to imply that fixed response testing is trivial or incapable of being used to test at all stages at tertiary level.

Alex Johnstone

formerly of
The Centre for Science Education
University of Glasgow

December 2003

Contents

SECTION 1: TALKING THE SAME LANGUAGE

As in every area of human endeavour, assessment has its own technical language and we need to become familiar with at least some of it. The problem with technical language is that old and familiar words take on a special meaning without informing the reader that this has happened.

Assessment as a Measuring Device

It has to be admitted that assessment, no matter how we disguise it, is a measuring device to provide information about progress and attainment.

There are four important considerations which must be applied to such a device. It must be valid, reliable, humane and economical in terms of time and cost. Let us take each of these ideas in turn and clarify their meaning.

1. **Valid** is a word capable of many meanings in common speech. It can apply to an argument, to conclusions and to railway tickets! It can range in meaning from "legitimate" through "true" to "current". Here is our first example of the problem of definition mentioned above.

> ***In the context of assessment VALID means that the test instrument measures what it sets out to measure.***

This seems to be utterly obvious and naive and yet many questions are set which are invalid. A good check for an invalid question is to ask if the question could be answered by someone who had not done the course. An example is given below which was supposed to test biological knowledge, but could be answered by anyone with a modicum of common sense.

The question shows a picture of two boys standing side by side and clearly they are of the same height. It also says that the boys are in the same class at school. It then goes on as follows:

> **In which one of the following respects are the boys most likely to differ?**
>
> A. Height; B. Age; C. Rate of heart beat; D. Number of toes.
>
> Common sense arrives at C by elimination of A, B and D.

This question, along with a number of others like it, appeared in a national biology examination! The setters were pleased with the high scores in the test and deluded themselves into thinking that the biology course was a success. Validity is all important.

In the assessment of practical skills it is possible to set very invalid questions without realising it. If a car mechanic is taught that a certain fault will give certain symptoms and if he is always tested in the "fault to symptoms" direction, there is no guarantee that he will be competent in operating in the real world where the direction is "symptom to fault". He could be a prize-winning student but a useless practitioner.

And so **valid** means that the test measures what the setter intended it to measure in terms of knowledge, understanding and skills. *If a test is invalid, everything which follows from it is useless*. No amount of skilled marking or statistical manipulation will make the test of any value.

However, assuming that our assessment tools are valid, they have a second hurdle to surmount. Are they reliable?

2. A question or a test is **reliable** when it produces similar results with the same students or with a similar group of students on the same course.

Strictly speaking, the same test should give the same score for an individual on two occasions close together, but this is a state of perfection rarely achieved. A student prepared for a test on one occasion will not necessarily perform as well on another occasion. Intervening experience or common forgetfulness will play a part in altering performance.

Where reliability is more likely to work is when the same test is applied to two large groups of similar students at the same stage in the same course. The average scores of the two groups could be similar and the spread of scores could be comparable. It is even possible that the average score on individual questions on the two occasions might be almost the same. This is particularly obvious when national examination boards, working with samples in the thousands, use the same questions in successive years. If the assessment is seen as a measuring instrument, it ought to be reliable. No one would use a measuring device, such as a thermometer, which gave different readings for the boiling point of water on different occasions.

However, a test or a question may be **reliable** and yet not be **valid** . The car mechanic's test mentioned above may give reproducible results which are measuring the wrong things. One can find *written* tests of laboratory skills which are reliable, but which are invalid because those skills may be tested only by *observing* them at the bench.

There are statistical tests of reliability which are arithmetically sound, but which may not be very useful if there are any doubts about validity. Common sense in looking at average scores and distributions is probably just as useful as calculating reliability coefficients which may appear to be erudite, but hardly worth the effort.

3. Our third consideration is **humanity**. Some tests in the past were gruelling, daunting and were sustained over days or weeks. Examples of these were the succession of papers which were (and sometimes still are) applied as "finals" at the end of a degree or of a high school career.

It is statistically true that the more questions asked, the more likely is the assessment to be a true reflection of the student's overall knowledge, but we have to stop some time! Flesh, blood and brain have limits. Exhaustion is not the best physical or mental state to do well in any assessment.

Maybe the reader does not indulge in such barbarism with students any more. Periodic, continuous assessment is, on the face of it, much more humane. A set of

progress and attainment snapshots may be preferable to the big set portrait at the end in giving a true picture of a student. However, a decision has to be made about how frequently such snapshots should be taken. Essays, tests, quizzes and exams can be so frequent and so time consuming that a student is in a permanent state of pressure and anxiety rushing from one deadline to another.

It must be borne in mind that no course is an island. The assessment in your course may be well constructed and humane, but students are simultaneously doing several courses and the cumulative assessment burden may be overpowering. This is particularly so in the new trends to modularise courses which have formative tests throughout and a summative examination at the end. As many as six modules may be running in parallel and teachers seldom confer with their colleagues to harmonise the assessment procedures.

4. The last of our considerations in this section is **economy**.

There has to be a balance between the time and resources spent on teaching and on assessment. There is no simple formula to determine what this should be, but there are some facts which help in making a decision about the types of assessment to be used. The form of the assessment is related to the cost in administering it.

If we consider the extremes of assessment typified by essays or long answer questions on the one hand, and multiple-choice or other fixed-response-questions on the other, a comparison of costs in terms of time and money can be made.

The setting of essay-type questions can be fairly easy, consisting of a sentence or two of a quotation and some instruction. It is almost possible to write such a question on the back of the ticket while travelling by bus! They may not be good questions by the standards we shall discuss in later sections, but they are common enough and they are cheap to set.

As far as responding to such questions is concerned, they take a lot of time. Exams typically last two or three hours and students have to write furiously to cope. However, the cost in teacher effort comes at the marking or grading stage. Hours of tedium, pints of sweat and gallons of midnight oil are expended in this exercise and, if costed realistically, this is frightening. Administrators, of course, hide this cost behind the good-will of the long-suffering teacher, but a price has to be paid in terms of neglect of teaching by jaded teachers.

At the other extreme, we have the fixed-response questions, typified by multiple-choice. Let us treat them in the same way as we have done for the extended-answer questions. Contrary to popular belief, they are fiendishly difficult to set. As we shall see in later sections they need a great deal of skill and time to prepare. It is not just a matter of one correct answer and three other absurd options. To make these questions work, all the options must be plausible and it is not easy to find them.

In terms of administration, the tests can be fairly brief; typically about one hour.

The grading can mercifully be handed over to mechanical devices which "read" the students' responses, "recognise" the acceptable ones, add up the scores and even do some neat statistics before spewing the results out on metres of paper.
In summary we have a situation like this (Table 1.1)

	Essay-Type	Fixed Response
Setting	**Fairly easy and cheap**	**Exceedingly difficult and expensive**
Responding	**Several hours**	**About one hour**
Marking	**Crushingly expensive**	**Fairly cheap if the equipment is available**

Table 1.1

As class sizes increase, there is a tendency to move to the fixed-response type of test to ease the marking burden, which is the obvious weakness in essay-type. However, the effort needed to set the fixed-response test is often grossly underestimated. The cost of this can be off-set only by "banking" used questions for use at a later date. This is a large task in that a bank of about 1000 questions is needed to set a series of 30-question papers and avoid undue repetition!!

Neither type of test can adequately assess the variety of abilities one might wish, and so a blend of both is necessary. This blend also tends to off-set the high cost of setting fixed-response papers against the high cost of marking essay tests .

It is time to sum up the substance of this section so far. We have been looking at four characteristics of assessment which must be borne in mind.

1. **Validity:** a test must measure what it sets out to test. If this condition is not met, everything else is a waste of effort.

2. **Reliability:** like any other measuring instrument, a test must give reproducible results on different occasions or with comparable groups of students.

3. **Humanity:** consideration must be given to the load which assessment places on the individual student, bearing in mind the other courses involved besides your own.

4. **Economy:** no form of assessment is cheap, but the costs lie in different places according to the type of test.

Comparing what with what?

Two other examples of assessment vocabulary now have to be addressed. There has been much discussion in recent years about what appear to be two distinctly different approaches to assessment; **norm-referencing** and **criterion-referencing** (with a sub-plot of **grade-related criterion-referencing**). What is all this about?

In any class one might expect a few high flyers and a few very poor students while the majority will be of "average" ability. A plot of the number of students gaining a given score would look something like Figure 1.1.

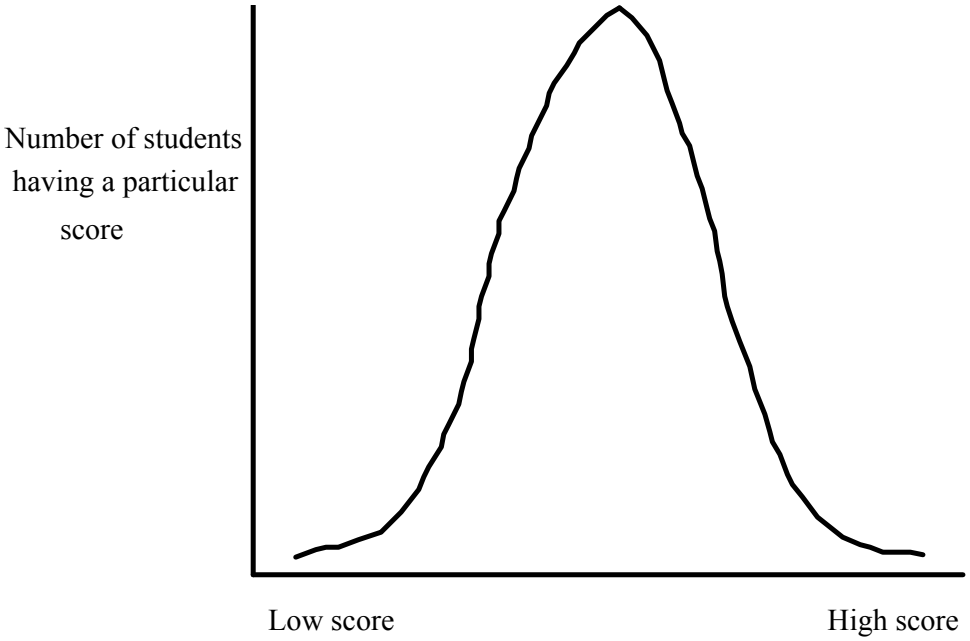


Figure 1.1

This is rather idealised, but is an approximation to reality. Even with very large classes, in the hundreds or even thousands, such a perfect curve is rarely obtained. However, for the sake of our discussion, let us assume such a "normal distribution". If a test sets out to discriminate between students to produce an order of merit (or rank order), a curve somewhat similar to Figure 1.1 is produced.

There now comes the problem of deciding on a number of boundaries. What score (or mark) is to be considered a *pass mark*? Where is the boundary between *pass* and *merit* or between *merit* and *distinction*? We have to make subjective judgments about the definition of these norms. It is not unusual to find the pass mark set mindlessly at 40% or 50% because it has always been so. The other boundaries may be set also by custom or by some dictat from "on high".

Another approach may be to follow some institutional policy which says that "not more than 5% of students can fail". If this is the case, you have to set your pass mark at a level which allows 95% of the students to pass. This is represented in Figure 1.2.

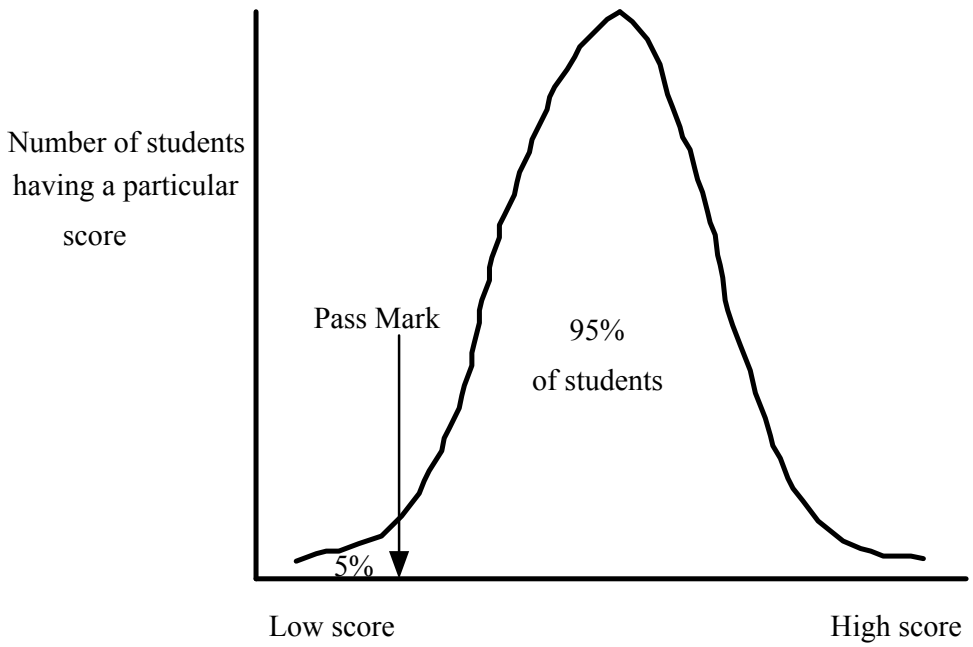


Figure 1.2

In the light of this decision, the other boundaries may be set.

Yet another approach is again applied institutionally where departments are instructed to give the top 5% of the students an A grade, the bottom 5% an E grade. The middle 50% get a C and the rest get either a B or a D grade.

All of this manipulation is called **norm-referencing**. It is assuming that there is an underlying distribution of marks as in Figure 1.1 and that students are divided by norms based upon it.

Some critics of this operation maintain that this is really making students compare themselves with each other and so introduce some invidious social division. In practice students pit themselves individually against the test and are pleased to pass, delighted to get merit and ecstatic if they attain the heights of distinction. One only has to observe students standing in front of a notice board displaying exam results to see this in action. The blend of tears, pleasure and elation is there. Students are fairly realistic about their performance. A fairly weak student will be ecstatic about a pass grade, while a high flyer will be disappointed with only a merit and shed some tears.

Most readers will have experienced norm-referencing and may be applying it to their students. This will continue as long as degrees are awarded by class (1st, 2.1, 2.2, 3rd, etc) and high school results are published as letter grades and entrance to higher education is conditional on such grades.

There is a serious problem with norm-referencing which has nothing to do with comparing student with student, but has everything to do with comparing exam with exam. Since each exam is usually freshly set for the occasion and judged against

notional (or even mythical) ideas of "**standard**", there is no guarantee that the distribution curve on one occasion is even remotely related to another generated on another occasion. A score of 50% on one test only has meaning for that test and probably cannot be equated with 50% on another test. Students and teachers have touching, but severely misplaced, faith in the value of 50% as a dividing line between pass and fail.

As we proceed through this guide we shall begin to see ways of adjusting for this and removing some, but not all, of the anomalies of norm-referencing.

Now let us turn our attention to the other end of the argument and look at **criterion-referencing**. This way of looking at assessment is not new; in fact it was the earliest of all. However, it has reappeared clad in a cloak of reasonableness and modernity, and has been used to bolster the *aims, objectives and outcomes* movement and has been used extensively in many national examinations. How does it differ from norm-referencing?

Let us have an imaginary trip back to some mediaeval swordsmith teaching his apprentice to forge a sword blade by some technique which produces a superior and much-sought-after sword. He teaches the lad how to select the iron, to heat it to the correct temperature, to hammer it in a certain way, to quench it and eventually to sharpen it to razor quality. He lets the lad do this over and over again, correcting, encouraging, cursing and praising until the lad either masters the craft or fails to grasp the skill and is relegated to forging nails!

The decision on the part of the craftsman to admit the lad's mastery or lack of it, is **criterion-referencing**. There is no acceptable half-mastery: either he can do it or he cannot. The master may extend the apprenticeship and give the lad further opportunity to gain mastery, but the master has a clear standard in mind by which he can declare the lad's success or failure.

Most of us will have experienced criterion-referencing in activities such as driving tests, music tests of instrumental skills, attainment of golf handicaps, or admission to membership of a learned society. In all of these cases, you succeed or you don't. There is no "almost" or "nearly", because there is a strictly (or loosely) defined standard or criterion which must be attained. The candidate is attempting to jump a hurdle of fixed height. It does not matter if you clear it with ease or just scrape over it, you have attained the criterion.

The reader can see how this ties in well with aims, objectives and outcomes. We now appear to have a very tight, objective way of assessing. General aims are set and the objectives are then written in such a way as to be unequivocal and universally agreed. They have to be couched in operational terms so that their attainment can be decided on a yes/no basis. Let us look at an example of this refinement process.

Someone is taking up rifle shooting and wishes to become a recognised marksman. Let us try to set up an objective which could be used as the criterion "hurdle".

> **"A marksman should be able to fire a rifle accurately."**

A nice idea, but it will not do. How will the examiners recognise "accurately" and what kind of rifle will be used? Let us try again.

"The candidate must hit the bull's eye nine times out of ten."

Surely that defines the word "accurately", but not at all! How far is the target away; is the candidate lying or standing; does the rifle have a telescopic sight; is it a small or large bore rifle and is the wind in the right direction?

You can see that to specify our objective in a way that would give an unambiguous criterion would be very detailed, very long and extremely tedious to write. However, such criteria can be written and provide the hurdle of correct height. But who says it is "correct"? It is a value judgment which makes all the pseudo precision in framing the criterion begin to look a lot less objective. However, it might be possible to get the international body to agree what is "correct" and then the criterion can be used. This applies to rules for games, driving laws and regulations and many things we commonly accept as reasonable.

Suppose that you find norm-referencing a bit old-fashioned or socially undesirable and yet you feel that criterion-referencing would not sit well with some more subtle outcomes which you would like to measure. Is there a middle way? Is it possible to acknowledge excellence and yet be specific enough to ensure comparability and reproducibility?

One possibility might lie with **grade-related criteria**. This is a compromise adopted by some national examining boards to try to get the best of both extremes. It recognises that it may be possible, for any given piece of learning, to set a series of attainment hurdles and the student would be given a grade related to the highest hurdle cleared. This seems reasonable until we consider the problem of defining the detailed criteria, not only for the pass/fail boundary, but for each of the other grade interfaces. This is task enough for one part of a course, but for a whole course lasting weeks or even months, the list of detailed grade-related criteria may become well beyond reason.

Let us look at one set of grade-related criteria, applied to the task of drawing a graph, to make the point.

The lowest hurdle
Given a pair of labelled axes and with all the points plotted, the student should be able to draw a straight line through the points.

The next hurdle
Given the labelled axes and five points, the student should be able to plot a sixth point and draw the straight line through them.

The next hurdle
Given a table of data and a pair of labelled axes, the student should be able to plot the points and draw a straight line through them.

The highest hurdle
Given a table of data, the student should be able to draw and label the axes, plot the points and draw a straight line through them.

For a skill like this it might not be too unreasonable, but for the assessment of oral presentation, report writing or problem-solving skills, the method is unlikely to be very useful.

Before we leave this territory, let us take a last look at the extremes of norm and criterion-referencing. Are they fundamentally different? In the norm-referencing situation, an examination is made up of many hurdles of different difficulties. There are easy questions and difficult ones and, even within one question, there may be routine easy parts followed by an unseen "sting in the tail". For a student to do really well, it is necessary to jump all the low hurdles and a large number of the higher ones. A student who does moderately well jumps the majority of the low hurdles and some of the higher ones.

Criterion-referencing has similarities with norm-referencing. How does one decide upon a hurdle's height without some experience of what "it is reasonable to expect from a weak student" or of "what an average student should be able to do" or of "how high flyers should behave"? Consciously or unconsciously, the criteria have to be derived from norms however notional. Maybe the whole business is just one conglomerate viewed from two angles. Where the extremes differ is that in one case criteria are specified and in the other they are unstated and almost certainly vague. This does not leave us in any happy and tidy state but merely points us in the direction in which we admit that both extremes may have merit in different situations. Where we can specify, let us do so, but do not try to press everything into the same mould.

Before concluding this section let us take a look at another area of language and definition which is frequently referred to in assessment: *the complexity of educational objectives.* This area provides us with a useful framework to help us with the **validity** of our testing.

The originator of much of this thinking was Benjamin Bloom who published a Taxonomy of Educational Objectives [1] in which he drew attention to the fact that educational objectives were not all of the same character, but placed different demands upon students.

What follows is a free adaptation, by the author, of his ideas applied to assessment. Objectives are placed on six levels and Bloom suggests that they form a hierarchy of complexity, but not necessarily a parallel hierarchy of difficulty.

The levels are normally called knowledge, comprehension, application, analysis, synthesis and evaluation.

Let us define (or describe) each of these six levels and what questions based upon them demand of students.

Knowledge	questions ask students to give back what they have been given by *recalling, defining, recognising and identifying*.
Comprehension	questions ask students to use knowledge in a familiar way to *describe, compare or illustrate.*
Application	questions expect students to use knowledge in unfamiliar situations by *applying, solving, classifying.*
Analysis	as its name suggests involves *concluding, inferring, deducing and distinguishing.*
Synthesis	requires the students to *construct, develop, predict and solve.*
Evaluation	is the weighing-up process in which students have to *argue, appraise, judge and decide.*

This provides us with a useful language for discussing the nature of questions and their demands on students. There is a temptation to confuse the degree of complexity of these levels with difficulty, but it is very possible to have very difficult knowledge questions and relatively easy analysis questions. The other pitfall is to try to use more than one level in one question. For example, if a student has to recall some knowledge and then use it in some way such as in application, the quality of the answer depends upon both levels. Because the student fails to recall something correctly it may be impossible for her to show her application skill. At which level the failure has taken place is often hard for the marker to discern. This is particularly so in fixed-response questions which we shall visit later. The basic rule is that if you want to test the skills of evaluation, provide the knowledge for the student to evaluate. In this way you separate the levels and test a skill with some confidence.

The levels, on a cursory reading, appear to be discrete and watertight, but there is no guarantee that what is application for one student is not just comprehension or even knowledge for another student who has read more.

Many users of Bloom's work lay great store by the Taxonomy being hierarchical. This has the unfortunate effect of relegating knowledge and recall to a lowly status and inflating the value of the others. A moment's thought should reveal the weakness of this. Without knowledge and the ability of recall, it becomes impossible for us to function. We depend upon a store of knowledge and know-how because without it we could not carry out the other five desirable skills. It is unwise for educators to relegate the knowledge base to some inferior level and glorify the others. We need them all and our assessment must acknowledge that all levels exist subtended by knowledge and know-how.

Perhaps a more useful way to think of Bloom's Taxonomy is shown in Figure 1.3.

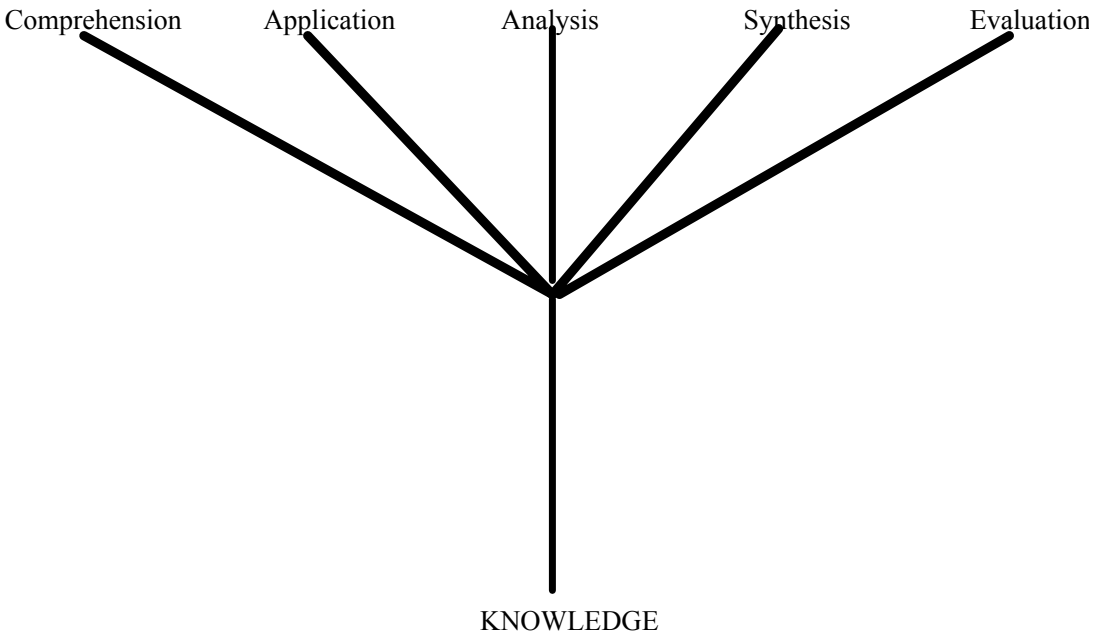


Figure 1.3

The extent to which examiners use this range of question types may vary from one discipline to another, but there would seem to be a case for use of all of them in different proportions.

An analysis, made in the 1970's, of the range used to test the sciences at university level, revealed that between 80-90% of the questioning was confined to knowledge and comprehension, and scant attention was paid to the others. Things may not have changed much since then! It may be necessary to have some departmental policy which indicates the proportions in which these skills are to be examined. We shall return to this when we consider test design in Section 5.

We have now set out the vocabulary which will be used throughout this guide. Other terms will be added as required, but the fundamental ideas elaborated here will underpin much of our thinking.

SECTION 2: CONVENTIONAL QUESTION TYPES

Before we begin to explore the field of objective testing, we need to clear up any potential misunderstanding. The word *objective* does not have its normal English meaning. These questions are objective ***only in their scoring or marking***. As with all forms of assessment, the objectives of the course are subjectively chosen and the questions are subjectively written to fit these objectives. In other words, professional judgment, rather than any strict objectivity prevails. The students then interact with the test materials, subjectively interpreting their meaning and responding to a fixed menu of choices. The computer, or the teacher, then scores these efforts against a subjectively determined answer scheme, but in an *objective* way. There is no room for the computer to interpret shades of meaning, or for partially correct answers to be recognised: there is an objective right/wrong decision made. Once the results are distilled out, subjective decisions are made about pass/fail, pass/merit and merit/distinction boundaries. The reader will now appreciate the tenuous nature of the objectivity of so-called objective assessment methods. However, they do differ from the more subjective methods (such as essays) in one very fundamental way: the use of language.

In subjective tests the language of the question, although important, allows for considerable interpretation. The language load is then placed upon the student to write an answer in some cogent, well argued and well presented way. The objective tests move the language onus to the question and its interpretation, while the response from the student may require no more than a tick in a box. Since the responses are fixed by the examiner, the student needs only to make a choice. *To make all the responses plausible, needs a subtlety of language on the part of the teacher and a sophistication of interpretation by the student.* Any idea that, in the face of students' alleged inability to use English, objective tests provide a "language free" solution, is just not so. If the language demand of objective questions is not high, the chances are that the fixed-responses are trivial. To examine at the variety of Bloom levels (Section 1), subtlety of language is almost always necessary, to get to the fine shades of meaning which separate the options.

It is also worth reminding ourselves that, although objective questions are said to be good for testing **recall,** they usually test **recognition**, which is **not** the same. This is the inevitable outcome of a process in which students are presented with options and have to make a choice. If you are struggling to recall the name of a colleague, the process is made simple if you are presented with a list of names including that of the colleague, to aid *recognition.*

Objective testing has a place in assessment, but only a place. To conduct all assessment by this method is not advisable. To cater for the range of student abilities, of testable objectives and student maturity, a battery of assessment tools is necessary. The most intellectually mature students generally hate objective (or fixed-response) testing because they need room to expand and show their independence of thought. There is a temptation (if not an incitement from "above") to adopt objective testing to cope with the rise in student population. While it does provide a relatively easy way to

process large numbers, it is by no means a panacea. It is subject to the constant, beguiling danger of *pseudo reliability* at the possible expense of *validity* (Section 1). If a test is invalid, no amount of computer marking and statistical manipulation can make the results meaningful.

This section will be devoted to helping you to recognise and avoid the pitfalls of objective methods, while helping you to gain the skills necessary to produce useful test materials. There is probably a place for this type of testing in all disciplines and so these skills are worth acquiring.

Types of fixed-response (objective) questions.

By far the commonest type is ***multiple-choice*** with all its variants. The classical multiple-choice question takes the form of a statement or question (called the *stem*) followed by a number of options (Figure 2.1).

STEM

A.....

B.....

C.....

D.....

Figure 2.1

The stem can be in the form of a partial sentence with each of the options grammatically completing that sentence.

Here is a fairly cynical example.

Students prefer multiple-choice questions to essays or short-answer questions because

A. you don't have to know anything to pass.
B. one Biro refill lasts a whole module.
C. it's like the football pools, anyone can be lucky.
D. you can finish and go home in twenty minutes.

The "correct" or "best" option is called the ***key*** and the "incorrect" options are called ***distracters***.

However, this example is really a poor specimen in that it does not ask a question. Must students select one option or can they choose more than one? Are all the options plausible or do they "put words into students' mouths" giving no room for anything but a cynical response?

We shall return later in this section to analyse questions in much more detail to see their strengths and weaknesses.

Another form of multiple-choice is the ***Classification Set***. This is useful for testing the ability of students to categorise data or objects (Figure 2.2).

A. Class 1. B. Class 2. C. Class 3. D. Class 4.

To which one of the above classes does each of the following items belong?

(i) .....
(ii) .....
(iii) .....
(iv) .....
(v) .....
(vi) .....

Figure 2.2

Categorising each item is effectively answering one question and so this would give a "nest" of six questions.

Yet another variation on multiple choice is the ***Situation Set.***

The stem is usually fairly lengthy taking the form of data, or a passage to be analysed or the description of a situation or even a diagram. A set of questions is then generated based upon the same stem (Figure 2.3).

DATA OR SITUATION		
Q.1	**Q.2**	**Q.3**
A	**A**	**A**
B	**B**	**B**
C	**C**	**C**
D	**D**	**D**

Figure 2.3

This type of question is economical of stem information in that one stem is rich enough to be used as the basis for several questions. Do not be misled into thinking that these are easy questions to set. Experience tells us that we have to scrap much of what we write either for content reasons or technical difficulties. Later in this section, under the heading *shredding* we shall look at this in detail.

The major drawback in classical multiple-choice questions is that there is no way of telling why students make the choices they do. They can, of course, guess blindly and setters try to offset this by increasing the number of options or by penalising students for wrong answers by deducting marks. It is almost certain that neither of these measures is effective. Increasing the number of options poses a problem for the setter. It is often difficult enough to find a key and three plausible distracters without having to find additional distracters. The extra ones are often so feeble as to distract nobody. In some scientific situations it is hard to find a key and even three distracters. A despairing example, which is genuine, is as follows.

Two measurements X and Y are being compared.

Which one of the following statements is correct?

A. X = Y;
B. X is greater than Y;
C. X is less than Y;
D. None of these.

The absurdity of squeezing out distracters to fit some prestated pattern, of one key and three or four distracters, is easy to see. It gets even worse as the number of distracters is increased!

It is instructive to ask students to answer multiple-choice questions out loud. Work described by Handy [2] tells of sitting with students as they attempted multiple-choice questions and listening to their reasoning. They adopted two main strategies. One was to scan the options for some key word that had been used in class and choose the option containing it. The other was to eliminate what they deemed to be weak distracters. If only one option was left, that was chosen as the correct answer. If two options were left, there was a mental "flip of the coin" and a choice was made. No cases were reported of a blind one-in-four guess; but guessing, where it occurred, was one-in-two after elimination of distracters.

There seems little point in increasing the number of distracters of dubious quality merely to have the students making a one-in-two guess in any case. If blind guessing was occurring, an increase in distracters gives little benefit in terms of eliminating guessing. Four options would give a guessing factor of 25%; five options would give 20%; six would give 17%. These gains are not worth the effort of finding convincing distracters.

Penalty marking, for example deducting 1/3 of a mark for a wrong answer, may deter guessing, but as suggested above, blind guessing may not be the problem. One has to ask if an "educated" guess is something to be discouraged. Professional people use educated guesses in all fields of activity. But we come round once more to the knotty problem of not knowing why students make the choices they do. We cannot tell if a guess is "educated" or not! Tamir [3] has shown that, of students who choose the "correct" answer, 30% of them do so for wrong reasons!

If the purpose of a test of this kind is to provide an order of merit from a class, it does not need too much statistical skill to see that scores generated by giving one point for a correct answer and none for any other, are in substantially the same order as those produced by giving one point for a correct answer and deducting 1/3 for each wrong one. The numerical scores are different, but their rank order is almost the same. (Rank order correlations of the order of 0.98 are not uncommon). There seems little point in the extra arithmetical manipulation to calculate negative marking.

The knotty problem of the reasons for choice has not gone away yet. Does a student make the correct choice for the wrong reason, or perhaps worse, does a student make a wrong choice for a good reason? Someone who has read more, or thought more, may see merit in what the teacher planned to be a distracter, but there is no way for that student to let the teacher know. In an open-ended question, that possibility exists. Able students find that having words put into their mouths in the fixed-response situation is frustrating.

Some research findings about multiple-choice questions.

In a recent study [4] standard multiple-choice questions were being offered to biology undergraduates to help in their preparation for an examination. Each student was offered a test on the computer and then received helpful feed-back.

However, each question in the test was randomly offered in one of three versions. Each version had the correct answer fixed in the same position, but the distracters were rearranged around it. Not one word or comma was changed in these three

versions; only the **positions** of the distracters were changed. An example is given below.

Anatomic structures that show similar function but dissimilar embryonic and evolutionary background are said to be

Version 1	Version 2	Version 3
a. primitive	a. primitive	a. homologous
b. polyphyletic	b. homologous	b. polyphyletic
c. homologous	c. polyphyletic	c. primitive
d. analogous *	d. analogous *	d. analogous *
e. monophyletic	e. monophyletic	e. monophyletic

* = key (correct answer)

Student responses to each version of each question were recorded by the computer. At least 150 students attempted each version giving a total sample of about 500 students.

For the thirty question test, the results are shown in Table 2.1. The numbers shown are the fraction of each sample of students who answered the question correctly (facility value).

	Facility Value				Facility Value		
Item Number	First Version	Second Version	Third Version	Item Number	First Version	Second Version	Third Version
1	0.44	0.61	0.58	16	0.57	0.64	0.55
2	0.63	0.80	0.78	17	0.96	0.98	0.97
3	0.85	0.94	0.90	18	0.65	0.73	0.68
4	0.44	0.46	0.44	19	0.71	0.66	0.68
5	0.62	0.57	0.59	20	0.74	0.52	0.57
6	0.83	0.86	0.79	21	0.93	0.89	0.88
7	0.58	0.58	0.59	22	0.78	0.77	0.80
8	0.56	0.62	0.46	23	0.73	0.71	0.69
9	0.44	0.32	0.28	24	0.81	0.78	0,80
10	0.27	0.33	0.31	25	0.60	0.64	0.55
11	0.54	0.46	0.49	26	0.36	0.30	0.45
12	0.58	0.56	0.57	27	0.51	0.50	0.56
13	0.81	0.75	0.93	28	0.67	0.61	0.55
14	0.79	0.70	0.79	29	0.76	0.84	0.86
15	0.59	0.59	0.47	30	0.49	0.51	0.49

Table 2.1

Let us look at the results for Question 1. In version 1, 44% of the students chose the correct answer; in version 2, 61% succeeded while in version 3, 58% got it correct. Remember that **not one word** was changed between the versions. Only the position of

the distracters changed. How then can we explain that essentially the "same" question in biological terms, when applied to very similar groups of students (randomly chosen by the computer), gave such different results? One would expect that if we were testing biology only, the success rate in all three versions would be very similar. There are some questions where this is so; for example, Question 4 is very stable with results of 44%, 46% and 44%.

A glance across the table shows a range of questions from the very stable (4,7,12,17) to the highly unstable (1,2,8,20). The instability suggests that these questions are testing something more than just biology. If minor editorial changes in the order of options can make changes in results of up to 20%, we must be concerned about the validity of such questions. Are they testing what we set out to test?

If by chance a teacher had set Question 20 as version 1, he would have been pleased with a class success rate of 74%, but he could equally well have set it in one of the other versions and obtained more disappointing results. Research continues to find an explanation for this instability, but these results should give some sobering thoughts to teachers planning to use this kind of testing.

Other types of fixed-response questions

There are several other methods which tend to be used in specific situations and in a few disciplines. Let us look at some of these so that you can decide on their usefulness for you.

Gap-filling questions are well named. The "question" consists of a statement or a piece of continuous writing with gaps left for the students to fill in a word or short phrase. In this respect they are not strictly fixed-response-questions because students have to supply something rather than select from a menu. There are, however, gap-filling "questions" in which menus are provided. Some examples are given below.

Example 1

> **Fill the gaps in the following passage.**
>
> The capital city of France is .....which is situated on the river.....
> It has many famous churches such as .....
> Another famous cathedral city in France is .....

Example 2

> **Fill the gaps in the following passage by selecting and underlining <u>one</u> of the words in each (space).**
>
> A member of the ranunculus family of plants is a (daisy, buttercup, cowslip).
> Dandelions belong to the family of (composites, geraniums, labiates).

Example 3

Using only the words in the following list, fill the gaps in the passage.

Andes, Pyrenees, Urals, Poland, France, Switzerland, Italy, Mont Blanc, Matterhorn, Eiger, Snowdon, Ben Nevis, Scafell.

Between France and Spain the range of mountains is called the .....
The Alps are located in several countries such as ..... and .....
The highest mountain in Europe is .....
Compared with it, Britain's highest mountain ..... is quite small.

These questions are largely testing recall or recognition and they do not lend themselves to testing the range of skills in the Bloom levels.

Matching sets are not uncommon in the sciences. They also tend to test recognition or perhaps comprehension (Section 1).

Example.

Here are two lists. List 1 names compounds and List 2 are descriptions of properties. For each compound in List 1 choose an appropriate property from List 2.

List 1	List 2
Q. 1. Potassium chloride	A. Powerful oxidising agent
Q. 2. Sulphur trioxide	B. Solid of high melting point
Q. 3. Ammonia	C. Powerful reducing agent
Q. 4. Chlorine	D. Acidic gas
	E. Basic gas

Notice that List 2 must be longer than List 1 so that the final "question" cannot be answered just by elimination of the first three.

True/false is really a very restricted multiple-choice. Ideally, in a multiple-choice question, each option has to be weighed up and be declared true or false. This is what was described earlier in Handy's work when students eliminated some options as false, before deciding on the true response. The options, in this case, were in some way related to the same theme to make them plausible.

However, in the most basic form of true/false testing, each "question", or perhaps a better name would be *item*, can stand alone and the student has to make a separate decision about each.

Example

1.	Bucharest is the capital of Hungary	T	F
2.	Juneau is the capital of Alaska	T	F
3.	York is east of the Pennines	T	F
4.	Hull is on the Mersey	T	F

Tests of this kind usually consist of one hundred or more items. As you can see, blind guessing, with the educational aid of a coin, would give students a score of half marks and so some way has to be found to discourage this. Devotees of this method of testing give one point for a correct answer, deduct one point for a wrong answer and award no points for an omission. Blind guessing or a total non-attempt would generate a score of zero!

This form of testing is common for medical students who, being intelligent young people, have devised dodges to "beat the system". The technique is to answer seriously the items about which they are sure, hazard very few guesses and ignore all the other items. In this way the positive score will outweigh the negative. This may be a good exercise in clever risk-taking strategies, but hardly a valid way of assessment. It is sobering to realise that such tests are in regular use even at the level of exams for surgeons and physicians. Admittedly, they are not the only test materials used, but sometimes they are used as a filter to select candidates for further testing.

A more sophisticated form of true/false testing occurs when a cluster of items share the same theme and then they become a multiple-choice question with the possibility of several correct answers.

A medical example is:

In venous leg ulceration

(a)	superficial thrombophlebitis is the main pathogenic risk factor.	T	F
(b)	abnormalities in fibrinolysis may be involved in the aetiology of the condition.	T	F
(c)	minimal pain is a typical clinical feature.	T	F
(d)	lipodermatosclerosis of the surrounding skin is characteristic.	T	F

In designing such a question care has to be taken to ensure that the answer to one item does not automatically eliminate the choice in another.

There are other forms of fixed-response questions such a Linked True/False, Venn diagrams and Structural Communication which we shall hold over to Section 4 to see how they help to reduce some of the weaknesses in the classical and commoner forms we have met in this section so far.

However, there are some variations on the classical questions which attempt to address the main drawback, the problem of trying to reveal some reasoning behind the

choices which students make. We shall look at these now before we proceed to the technical skills of writing fixed-response questions.

The first of these is called ***multiple-completion***. In questions of this type the shape is similar to that of multiple-choice, but one option or a cluster of options may be correct. The cluster has to be self-consistent, providing a number of facets of the same concept or supporting a statement with evidence. Guessing becomes almost impossible because the student is not told how many options to select, but has to look for some internal consistency among them. As the reader can anticipate, these questions are exceedingly difficult to set, and few published examples exist. A lesser version is, however, common. This becomes a multiple-choice item with the potential for more than one correct response, but the responses need not inform or support each other. The aim of trying to find student reasoning can easily become lost. Most versions of *multiple-completion* are embedded in a peculiar coding system which tries to press the response patterns into an A, B, C or D format to make them computer markable.

Here is an example.

S T E M

1. Option
2. Option
3. Option
4. Option

Choose A if option 1 only is correct

Choose B if options 1 and 3 only are correct

Choose C if options 2 and 4 only are correct

Choose D if option 4 only is correct

The laudable aim of tracing student reasoning is in danger of being completely lost in the complexity of the instructions. If a student makes a wrong choice of A, B, C or D the examiner does not know if the reasoning in the question was wrong or if the use of the code was faulty. This coding also cuts out the possibility of partial credit for partial knowledge.

Another anomaly occurs if a student is certain that option 3 is correct, but is uncertain about the others. The choice of B has to be made, implying that the student agrees with option 1, which may not be so. In these days of more sophisticated computing, it is perfectly possible to do without the code and allow the computer to scan any combination of student choices, thus restoring the potential of this type of question to examine more deeply than the conventional multiple-choice.

Another variation with potential for testing reasoning is called ***assertion-reason***.

These items do not look like multiple-choice and are in fact two true/false items linked (or not) by some reasoning. Here is the format.

STATEMENT 1	**STATEMENT 2**
If statement 1 is correct and statement 2 is false, choose	A
If statement 1 is false and statement 2 is correct, choose	B
If both statements are false choose	C
If both are correct, but statement 2 does not help to explain statement 1, choose	D
If both are correct and statement 2 helps to explain statement 1, choose	E

Once more the code gets in the way of what the item is trying to test. No provision is made for the student who recognises that both are true, but does not see any relationship between them. Modern computers could cope with this by the student interacting directly with the machine to reveal reasoning.

One other example of pursuing reasoning in fixed-response questions is called ***conceptual testing***. In this method, questions are set in pairs, the first asking about factual information and the second looking at the possible underlying reasoning or concepts. The first is specific and the second more general.

An example of such a pair might be:

Which one of the following flowers belongs to the labiate family?

A. Dandelion
B. Dead nettle
C. Buttercup
D. Cowslip

The reason for your choice was that

A. dandelions have many petals.
B. dead nettles have square stems.
C. buttercups have five petals.
D. cowslip flowers are carried on a common stalk.

Another example is:

Which one of the following metals would release hydrogen from cold dilute hydrochloric acid?

A. Copper
B. Gold
C. Magnesium
D. Silver

The reason for my choice is that

A. copper chloride is a common compound.
B. gold atoms are the largest.
C. magnesium is more reactive than hydrogen.
D. silver chloride is insoluble in water.

These questions, though well intentioned, present problems in setting and also in making sense of the responses.

In the first example, supposing the student rightly chooses B (the dead nettle). In the follow-up question, all the responses are factually correct, but the student is going to choose B again, just because he chose "dead nettle" in the first question and not necessarily because he recognises the importance of the square stem in identifying labiates or that dead nettles have square stems.

A similar criticism can be levelled at the second example when the correct answer to the first question is C (magnesium).

This question type does not solve the problem of tracing student reasoning. To be of any use, each initial question would have to be accompanied by a cluster of four follow-up questions only one of which would be answered by the student depending upon the choice in the initial question. This could be handled easily if the student was undergoing the test while sitting at a computer.

Depending upon the response to the initial question, the computer would present the student with another question which offered optional reasons for the original choice. Setting such questions would be time consuming and quite difficult, but feasible. However, the problem of the train of reasoning is not solved if none of the options offered is the *actual* reason for the student's initial choice.

Considerable efforts are being made in the United States to produce such "conceptual questions", but information about their usefulness is, as yet, scarce.

There are some more recent forms of fixed-response questions which we shall examine in Section 4, but let us now turn to the techniques for setting the fixed-response questions which have been outlined in this section, with emphasis on multiple-choice and its variants.

SECTION 3: CONSTRUCTING FIXED-RESPONSE QUESTIONS

It is worth re-emphasising that fixed-response questions are not easy to write. If you write a stem, decide on a correct answer (the KEY) and then dash off another three distracters, the chances are that the question will turn out to be disappointing.

The embryonic question: the conception

The first step is to get an idea for a question and refine it later. Some of the best ideas will arise while teaching, particularly for questions testing the whole range of the Bloom classification. When you are reading, be on the look-out for information which could provide the stem for a set of questions designed to test the ability to apply understanding to new situations (*Application* in Bloom's classification).

The best way to find distracters is to note the replies you get from students to discussion points in tutorials. These represent their real misunderstandings rather than ones you might invent for them. Keep a note pad handy to record these ideas at the end of the tutorial before you forget them.

There are two ways of making questions difficult. One is to give students a difficult problem or difficult information to handle in the stem. The other is to make fine distinctions between the distracters and the key. This will often depend upon subtlety of language when you play with adverbs like "usually", "generally", "predominantly", "slightly" and so on.

Having found an idea, write the question in full, leave it to mature for a few days and then read it again to see if it is as good as you thought. If the question survives, refine it until it is ready to go for the process of **shredding** which will be described later in this section. Make sure that you attach to the question the following information for the person who will assemble the whole test. Note the **key** (the correct answer), the **objective** you are trying to test and the syllabus **topic**. Consider the effect of the question on later learning, because students (and other teachers) will tend to use the questions as interpretations of the objectives of the course. Too many questions on any topic can give an enhanced impression of its importance.

Do not try to write too many questions at a stretch or you will soon find yourself becoming stale and mentally numb. Five or six questions of good quality are a full night's work for most people especially if you wish to test more than recall or recognition.

The last and most painful tip in the writing process is to be prepared to scrap much of what you write. It is easy to become too clever and too perverse in writing questions and the bright idea which came to you in the shower, or when walking the dog, may turn out to be beyond any reasonable expectation of your students' abilities. However, ideas are precious and should be filed away for use at a later time or for students at a higher level.

The embryonic question: the gestation

Having produced your questions you should now check them for some of the common faults which can so easily occur in them.

(i) Superficial clues

There may be something in the stem or the options which guides the student to the correct response for reasons unconnected with the objective you are trying to test. For example, if you have used an important technical word in your teaching and it appears in only one of the options, the students will select that option, not on the basis of understanding, but simply by word recognition.

(ii) Options of uneven length

An option which is much shorter or much longer than the others tends to draw undue attention to itself. A long option can be especially attractive if the question calls for an explanation, because the student assumes that the fuller the response, the better the explanation. This fact alone should make you wary of fixed-response questions when a technicality, like option length, can influence student response. If it is at all possible, make the options roughly the same length.

(iii) Negatives

It is important that, if there are any negatives in your question, you should highlight them in some way. In the heat of a test, it is only too easy for a student to fail to see a negative and give a false response. Even if the student notes the negative, it is more difficult to decide on the truth or falsity of a negative statement than it is of a positive one [5]. "**NOTS**" should be in bold type or in capitals. However, negatives are not confined to the word NOT, but take in words like "never", "unlikely", "smallest", "least", "false", "colourless", "inert" and many more.

Indeed, anything which *implies negativeness* should be emphasised in some way.

(iv) Double negatives

An extension of (iii) above brings us to the dreaded double negative. This can most easily occur when there is a negative (stated or implied) in the stem and then a negative, in some form, appears in an option. Double, or even triple negatives, make it exceedingly difficult for students to unpack, interpret and respond to such questions. As I pointed out earlier, the problems of language are at their most acute in the *interpretation* of fixed-response questions. Although the student response may be a tick in a box, the onus of language is still very much in evidence as the student tries *to make sense* of the question.

The Scottish clan from which my family springs has a motto, "Never unprepared". It would have been clearer to me, and everybody else, if it had been written as, "Always ready"!

Go over your questions and root out any hint of double negatives and, if possible, avoid any negatives at all. Some of my own researchers [5] have found improvements in student scores by as much as 15% when questions were presented in a positive form rather than in a negative one! In this case, a psychological artifact is getting in the way of valid assessment.

Here is a question to exemplify our thinking so far.

> **Which one of the following best describes the type of option which should NOT appear in a fixed-response question?**
>
> A. One which never uses negatives.
> B. One which is longer than the others.
> C. One which fails to confuse the student.
> D. One which avoids complex language.

It is a dreadful question, but it should make the point of the "double-think" which negatives impose on students in tests of this kind. Even "fails" and "avoids" have negative connotations!!

(v) Instructions to students

First of all make sure that you have asked a question or have given clear instructions to the students.

> **An instrument for the measurement of air pressure is**
>
> A. a barometer. B. an anemometer. C. a pedometer. D. a photometer.

This is not a question and would be improved by a new stem which might be:

> **Which ONE of the following is the instrument for measuring air pressure?**

There is now no doubt about what the student must do. You might have a question in which there is more than one correct answer and the students, used to one correct answer, need to be alerted to the possibility of multiple answers. This must be indicated in the stem by some device such as:

> **Which of the following instruments may be used to measure temperature? More than one correct answer is possible.**
>
> A. Pyrometer. B. Bolometer. C. Manometer. D. Odometer.

If the instructions are complicated, as we saw in the multiple completion and reason-assertion questions, it is a good idea to gather all the questions of this type together in the test paper to avoid having to repeat the rubric. This is not just an editorial device, but it helps the student if the format is not always changing from question to question.

A similar clustering under one rubric would accommodate the type of "non-question" which we criticised at the beginning of this section.

Such a rubric might be:

> **In the next section (in questions X and Y) choose the response which best completes the sentence in the stem.**
>
> **Question X -**
>
> A philatelist is one who collects
>
> A. matchboxes. B. postage stamps. C. signatures. D. old books.
>
> **Question Y -**
>
> Radioactivity is
>
> A. what is broadcast on the radio.
> B. the emanation of particles from an atom.
> C. the signal picked up by radar.
> D. the treatment for some diseases.

(vi) Grammatical Construction

To make the students' task clearer, it is important that all the options are in the same form and follow logically and grammatically from the stem.

In the examples we have just seen of sentence completion, the stem leads directly to the options with each option beginning with a small letter and finishing with a period, so that the stem plus **each** option reads as a complete sentence. Confusion ensues if this simple pattern is not followed.

Another grammatical effect which can unduly affect student response is seen when options are of different construction. One may be a complete sentence while others may be single words or short phrases. The complete sentence option generally attracts more attention than the others.

In all of these comments about grammar and question layout, we are once again confronted with problems about validity. If situations of the kind set out in this section can cause differences in student response, we must ask what we are really testing.

(vii) Self-cancelling responses

There is a temptation to ask questions about the "most" or the "least" and to offer the options in some well known order. This is particularly so in a discipline like chemistry where a group of elements (or their compounds) are set out as options in the familiar order of the Periodic Table. An example would read something like this.

> **Which one of these elements has the smallest atoms?**
>
> A. Lithium. B. Sodium. C. Potassium. D. Rubidium.

Since this is a well-recognised sequence, the answer can only be A or D. Distracters B and C are self-cancelling.

There is another sense in which options can be self-cancelling, which is seen in multiple-completion questions with their accompanying code.

There are four options labelled 1 - 4 and combinations of these give rise to a letter code

A. if only 1 is correct.
B. if only 1 and 3 are correct.
C. if only 2 and 4 are correct.
D. if only 4 is correct.

When the student looks at the options 1 - 4 and finds that 1 is a positive description of the behaviour of something and 3 is a negative description of the same thing, it is highly unlikely that option B will be chosen. This is not testing his knowledge, but his common sense. Students can be drilled, in advance of tests, to look out for such things and obtain respectable scores more by strategy than by knowledge and understanding.

(viii) Removing redundancy

This is a tidying-up operation, but it goes further than just editing.

This is best shown by an example. This is a poor question on a number of counts, but its form will illustrate the point of this section.

> **A flame test was carried out on the evaporation residue of sea water, when a strong yellow colour was seen. It can therefore be said**
>
> A. that sea water contains sodium chloride.
> B. that the sea water contains sodium atoms.
> C. that sea water contains sodium ions.
>
> **Which statement is correct?**

The question could be tidied up by removing the redundancy from the options and taking the repetition into the stem. The question would then read:

> **A flame test was carried out on the evaporation residue of sea water, when a strong yellow colour was seen. It can therefore be said with certainty that sea water contains**
>
> A. sodium chloride. B. sodium atoms. C. sodium ions.

Not only does it remove redundancy, but it focusses the students' attention to the main issues clearly presented in the options.

As a general rule, so structure the question as to take most of the verbiage into the stem and leave only the main issues for decision in the options.

(ix) Question length

It must be borne in mind that in a test fixed-response questions generally earn the same mark credit and so very wordy or complex questions are probably better in some other format such as structured, open-ended questions. This particularly applies to questions needing an amount of calculation for each option. The question makes a disproportionate demand on time and is not really suited for a fixed-response format.

If a large amount of data appears in the stem, a nest of questions based on that data might make it worthwhile (see Situation Sets, Section 2) and effectively award more marks for it.

(x) Test editing

This last section is really for the person whose task it is to compile the whole paper from questions submitted by colleagues. Each question may have been refined and polished into a gem, but stringing the necklace needs great care. The things to look out for are set out briefly below and we shall return to them later when we consider test construction in more detail.

(a) Make sure that a question set by one colleague does not contain the answer to a question set independently by another. Sometimes the information in the stem of one question is the answer to another.

(b) When tests are printed, make sure that the stem and the options are on the *same page.* The effort of turning pages or of going from the foot of one page to the top of the next facing page or, even worse, having a question on both sides of the same page, detracts from the students' ability to answer questions.

(c) Make sure that there is a "house-style" about the questions so that confusion between formats is avoided.

The embryonic question: the delivery

You have set your questions and have polished them in the light of this "gestation" section and they are "ready to go". But wait: there is one more process to undergo and it is the most painful of all. It is called **shredding.**

Shredding is the process in which a team of setters sit round to criticise (in the nicest way, of course) the work of colleagues. Your pet questions are about to be subjected to a refining process in which you will find it hard to believe that you were so blinkered. They will find ambiguities, inaccuracies, technical points and faults that you never saw. Comfort yourself with the thought that you will have the chance to do to others what they are about to do to you! This is a very necessary and professional stage in test preparation. If ambiguities are not caught now, students will find them, and the results you get from the test will be distorted. So, grin and bear it and put it down to professional training.

Suggested shredding sequence

Rather than glancing at questions to look for faults, it is helpful to have a check-list against which to evaluate each one. I offer a sequence of checks to which a yes/no answer is appropriate. This will enable the shredding team to carry out a systematic scrutiny during which a question's faults can be isolated and remedied or, if the faults are fatal, the question can be totally rejected. It may seem to be a tedious exercise to go through this check-list for each question, but its systematic use will pay off in well-crafted questions which can later be banked for reuse.

The check-list

1. Is the question factually correct?

2. Is the reasoning too complex?
   (a) Too many reasoning steps?
   (b) Calculations too complicated?

3. Is the question independent of the others in which a clue may be lurking?

4. Are all the options within the syllabus?

5. Is there more than one correct response?

6. Are all the distracters likely to function?
   Check for (a) implausible distracters
   (b) self-cancelling distracters
   (c) extremes such as "most" or "least".

7. Are the stem and the distracters properly stated?

   Check for
   (a) ambiguity
   (b) conciseness
   (c) grammatical consistency
   (d) punctuation
   (e) negatives properly indicated
   (f) relative lengths of responses
   (g) simplicity of language
   (h) double negatives (or implied negatives)

8. Does the question test the aims of the course in terms of its content and spirit?

9. Has the setter clearly indicated, for the test assembler, the key, the Bloom level and the likely degree of difficulty?

Questions in the Bank

I hope you are now convinced that setting fixed-response questions is no easy business, but there is a bonus. If your shredded questions are now used in a test and they function well, they can be "banked" for future use. This bank deposit is precious in that compiling papers in the future becomes easier, but one needs a very large bank deposit if papers are to be compiled to avoid frequent repetition and to meet the specification criteria which we shall discuss in Section 5. A rough rule-of-thumb is that the bank needs to be *at least* **ten** times larger than the number of questions you intend to set in a paper. A one hundred question paper ideally needs a bank of one thousand questions to sustain it, avoiding repetition, but giving flexibility.

It is very difficult for a department to set up such a large bank and so it is necessary to cooperate with other institutions to share banks. Some learned societies offer the services of a large bank, but not all the questions are necessarily suitable for your particular course. It is not possible, in practice, to avoid the chore of setting your own questions and supplementing them from a central bank.

There are some devices for multiplying successful questions by which you can swell your bank fairly painlessly. For example, if there is a good question testing the behaviour of copper, one can write parallel questions on nickel or cobalt. This "cloning" is easy to apply in the sciences and similar disciplines.

Banking statistics
The usefulness of a bank is greatly enhanced if you attach some extra information to each question before you bank it.

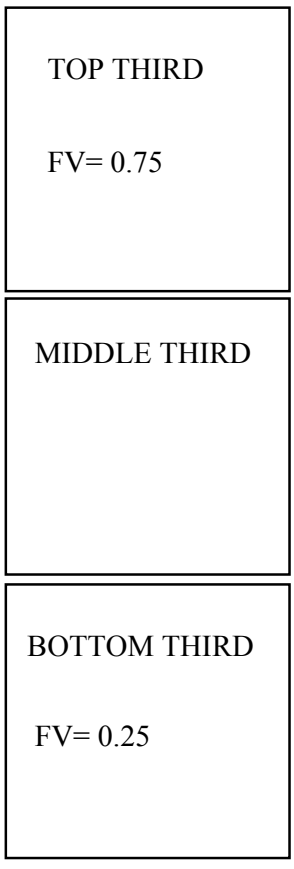
When a question has been used in a test, it has been subjected to the "ultimate shredding" by students. You will be able to decide if the question was easy or difficult by seeing what proportion of your students chose the correct answer (the **key**). That proportion expressed as a fraction (e.g. 0.6 or 0.2) is called the **facility value** (FV) of the question.

You will also be able to see if your distracters have functioned. If few students chose a distracter, it does not warrant the name and the question will have to be rewritten with a more attractive distracter. This distribution of student responses to the key and the distracters is useful information to help you to decide whether to bank an item or not. A rough guide is that each distracter should attract about 5% of the student responses if it is to be regarded as a genuine distracter.

One other piece of information which is useful is the extent to which the question discriminated between the able and less able students. This is really an issue for tests which are norm-referenced (Section 1) in which we are trying to separate students into an order of merit. However, if the test is criterion-referenced, the test is simply discriminating as pass or fail.

There is a simple but tedious arithmetical way of deciding if a question discriminates well or not. This is usually left to the computer, but some idea of how it works and what the values of a **discrimination index** (DI) mean, are shown below.

When the test is over and you have the scores for the whole class in order of merit, divide the list into a top third, a middle third and a bottom third *on the test as a whole* (Figure 3.1).



$$DI = 0.75 - 0.25$$
$$= 0.50$$

Figure 3.1

Now look at *a particular question* and work out the facility value on that question for the students in the top third of the class (i.e. what proportion of them got it correct). Do the same for the students in the bottom third of the class.

Suppose that 75% of the top third were successful (i.e. FV = 0.75) and 25% of the bottom third were successful (i.e. FV = 0.25) then the discrimination index would be the difference between 0.75 and 0.25, which is 0.50. This question would have discriminated very well. However, if the two facility values were almost equal, the discrimination would have been poor. This means that, on this particular question, the poorest students did as well as the best. This question made no contribution to the test as a whole as a discriminator. On this basis, a decision has to be made about the usefulness of this question and whether it should be banked.

In most computer programs for handling fixed-response questions, a more sophisticated discrimination is calculated (called point biserial correlation), but it leads to the same decisions being made about banking or rejection.

A very simple and visual way of handling discrimination and other useful factors will be described in Section 6.

SECTION 4: FIXED-RESPONSE QUESTIONS WITH A DIFFERENCE

As we have thought about objective, fixed-response questions there have been recurrent themes of unease.

When a student makes a choice in a fixed-response situation, is the correct choice being made for a right or a wrong reason? When a wrong choice is made, is it out of ignorance or for a very good reason? Few, if any, of the fixed-response types we have seen in the previous sections allows the examiner to ascertain this reasoning background. Assumptions are made that if a student selects a distracter and not the key, this necessarily indicates a certain lack of knowledge or betrays some specific confusion. However, if the differences between the key and the distracters are subtle (as they should be at tertiary level), there is some degree of "rightness" in all the distracters and students who have read more, or thought more, might choose a distracter instead of the key for very good reasons. Unfortunately, the nature of these questions does not provide a student with the opportunity to "tell all".

A second part of our unease stems from the fact that students who have moved on to the level of intellectual maturity, find fixed-response questions irksome, confining and frustrating and they often choose distracters rather than the key for very good reasons.

This sometimes shows as **negative discrimination** for some questions in which the facility value for the top third of the class is **less** than the facility value of the bottom third! The more discerning students have chosen distracters other than the key while the less insightful (or more naive) score well because they cannot see the other possibilities.

All of this presents real problems for the use of fixed-response questions unless other types can be found which provide the advantages of fixed-response (e.g. rapid and reliable marking) with a lessening of the disadvantages.

Our research has shown that if the same area of learning is assessed by open-ended, subjective methods and also assessed by objective, fixed-response methods, two orders of merit are generated for a given group of students. One might expect that, since the same knowledge and understanding is being assessed, the two orders of merit should be substantially the same for the same sample of students. The best student by one method should be the best by another method and so on down the line. But experimentally this turns out not to be so. If a correlation is worked out between the two orders of merit, it usually comes out at about 0.6. This figure turns up frequently in the research literature. For those not familiar with a numerical value for rank order correlation, a word of explanation may be necessary. A perfect match in order would result in a value of 1.0; a complete reversal of the order would give a value of -1.0. A completely random pair of orders would give a value of zero. The experimental value of 0.6 suggests that the two orders of merit have some similarity, but are by no means well matched.

This drives us to ask why the orders do not match. Much experimental work has been done to try to answer this question [6] and the most important factor to emerge is that, in scoring the open-ended questions, credit is given for partial knowledge or for wrong conclusions arrived at for good reasons. In the fixed-response situation, no such credit is given. This brings us back to the problems we raised at the beginning of this section about the lack of evidence for student reasoning.

There have been several ingenious attempts made to score multiple-choice questions to allow for partial knowledge. Some of these ask the students to rank all the responses in the question from the best to the worst. In other cases students are given a tick and two crosses and asked to use the crosses to label distracters they know to be wrong and the tick to choose the key. They get credit for throwing out the wrong, as well as for choosing the correct. These are obviously more difficult to score. The rank order produced when these devices are applied to multiple-choice tests and the rank order produced by an open-ended test, of the same scientific content, correlate to give a value of about 0.9; almost a perfect match. This underlines the importance of the examiner having the means of detecting and rewarding reasoning.

Thinking about newer forms of fixed-response questions has turned to giving credit for partial knowledge and for indications of reasoning paths. The remainder of this section will be devoted to examining three fixed-response formats which attempt to make allowances for the weaknesses of the conventional formats.

1. Interlinked True/False Questions

In Section 2, we saw examples of conventional True/False questions where each true/false decision stood alone and was independent of the questions before or after. In this interlinked format, each true/false decision has consequences for the next decision and so on along a chain. Let us take a stylised example (Figure 4.1).

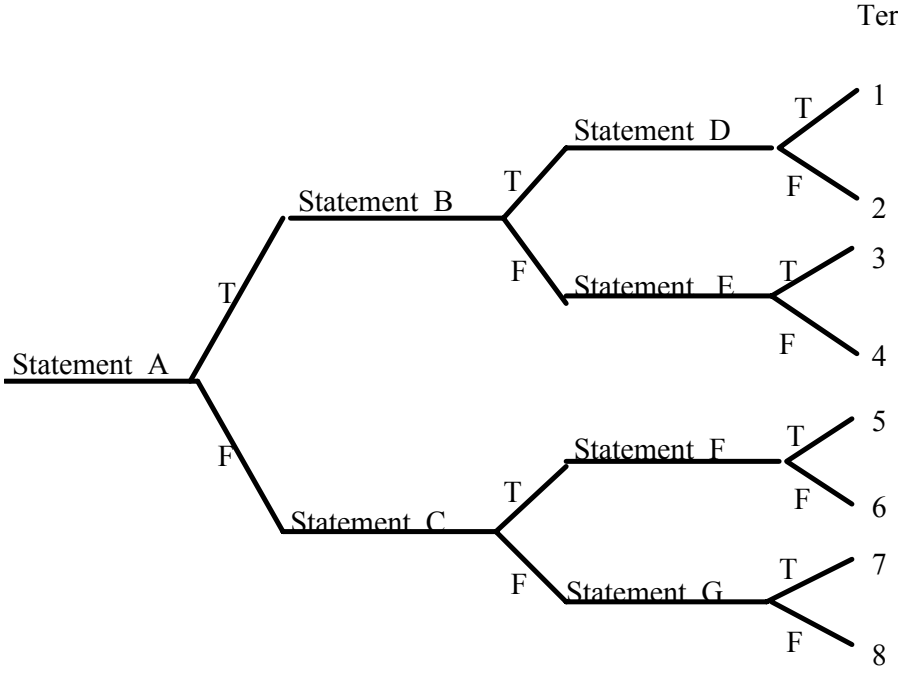


Figure 4.1

This is best done on a computer, but paper methods are possible.

Every student is presented with statement A and asked to pronounce it to be either true or false. If the decision is "true", the student is presented with statement B which is in some way a consequence of that decision.

If, however, the decision on the statement A is "false", the student is directed to statement C which is, in some way, a consequence of that decision.

This process continues from statement B to either D or E or continues from statement C to either statements F or G. Finally a decision is made about D (or E) or about F (or G).

Each student makes decisions on only three statements, each one a consequence of previous decisions. This brings the student to a terminus (numbered 1- 8), each one of which uniquely defines the route taken by the student. For example, a student arriving at terminus 3 has declared statement A to be "true", statement B to be "false" and statement E to be "true".

Let us suppose that the "best" answer is to arrive at terminus 4. (Students arriving here purely by guesswork, would have a one in eight chance of doing so.) Terminus 4 would get the best score, terminus 3 would get credit for two correct decisions on the way and each of the other termini could be given partial scores. Even students who arrived at termini 5 - 8 would get credit for correct decisions on the way, although their initial decision on statement A was wrong. Also, each terminus could carry diagnostic and remedial help for each student.

An example of a set of linked true/false statements is shown below (Figure 4.2).

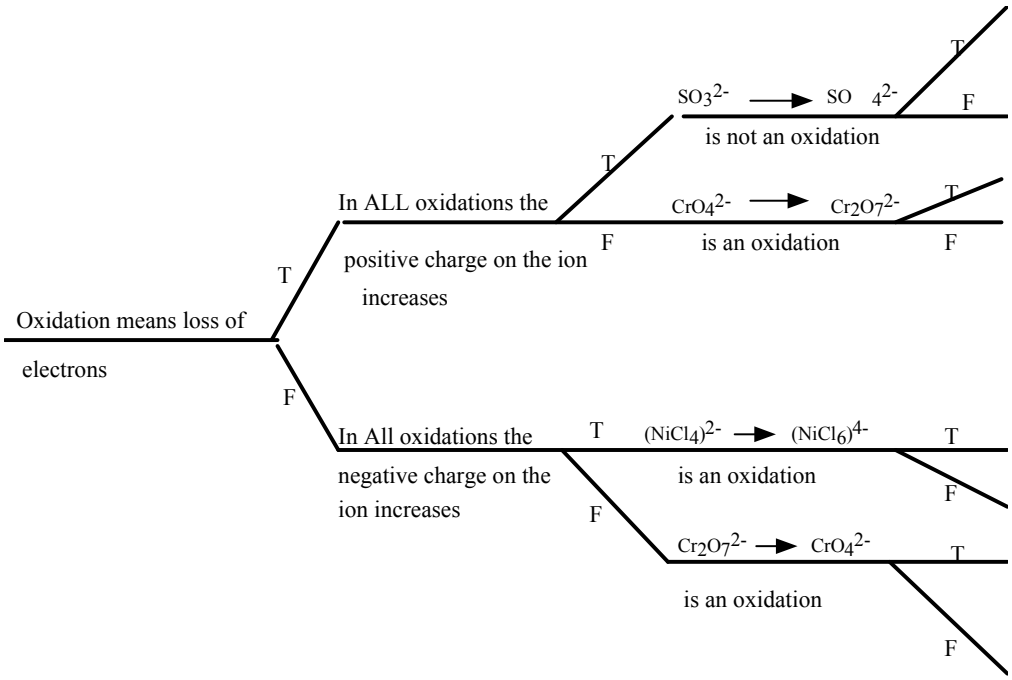


Figure4.2

This is attempting to explore some misapprehensions which are commonly held about oxidation. The best sequence would be: true, false and false, but other combinations would merit some credit.

This kind of question set is best administered by computer and then each choice-point (*or node*) can be made even more sophisticated. On the way through a true/false sequence, a student may realise that a wrong choice has been made further back and the possibility of back-tracking is needed. This is easily achieved by computer and the machine can keep a record of the student's choices and back-tracking to give the teacher information by which the student can be helped later. Programs are available into which a teacher can fit a set of statements with no programming skill and so construct a range of interlinked true/false statements.

A further use of such interlinked true/false questioning is to prepare students for a discussion in which to explore a series of decisions and follow their consequences. For example, medical students are presented with a situation which may take several sentences to describe and are then asked to make a decision (sometimes against the clock). Depending upon the decision, a consequence appears to which they have to react with another decision. This can lead to yet another consequence and so on. A set of wrong decisions might lead to the "patient's" death or a combination of rights and wrongs may require some drastic action to restore the situation. Similar decision exercises can be devised in the sciences.

Situations like this can be handled in groups to stimulate discussion or individually as a form of assessment.

2. Venn Diagrams

This is a simple, pictorial form of assessment which allows for degrees of "correctness" and is best used in situations which require the *ability to categorise* (Bloom *Application*). These diagrams are used in teaching mathematics and other subjects to take a logical approach to categories, sub-categories and shared categories. Some examples are given below (Figures 4.3, 4.4, 4.5).

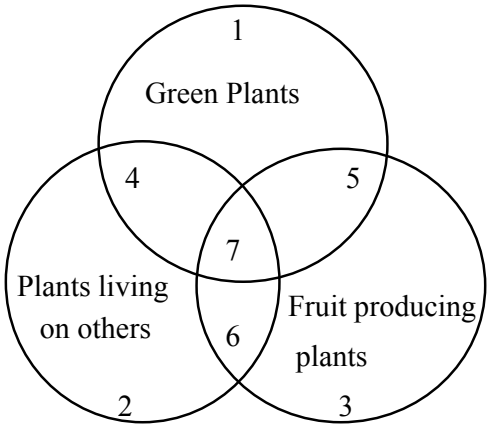


Figure 4.3

Into which area of this diagram do the following plants best fit? Indicate your choice by writing the area number against the plant name.

PLANT	AREA NUMBER
(a) holly	
(b) mistletoe	
(c) cherry	
(d) lichen	
(e) liverwort	

The best answer for mistletoe would be area 7 because the plant has green leaves, has berries and uses oak trees as its host from which it draws nourishment. However, a student may know that mistletoe has green leaves and white berries, but may not know that it is dependent on oaks. This student would choose area 5. The nature of the partial knowledge becomes obvious to the examiner. Similar evidence of partial or even wrong knowledge is made evident by the student's choice of area for each of the plants listed. Scoring can be weighted to take account of these choices.

Another example might be:

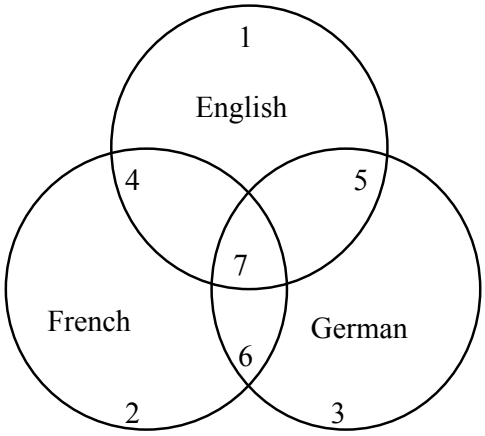


Figure 4.4

In some countries one or more languages may be officially spoken. Choose the area in the diagram which best represents the language(s) spoken in each of the following countries and enter its number opposite the name of the country.

COUNTRY	**AREA NUMBER**
(a) Switzerland	
(b) Canada	
(c) Monaco	
(d) Austria	

The Venn diagrams need not be a set of three intersecting circles. An example might be as in Figure 4.5.

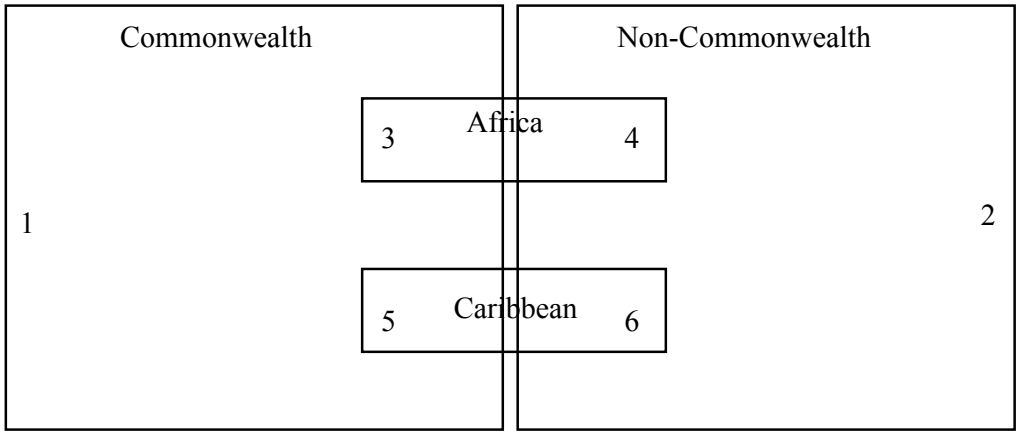


Figure 4.5

Into which area in the diagram do the following countries best fit. Indicate your choice by writing the area number against the name of each country.

COUNTRY	**AREA NUMBER**
(a) Nigeria	
(b) Barbados	
(c) Malaysia	
(d) Cuba	
(e) Algeria	
(f) Brazil	

These questions are easy to set and to mark and they give an indication of partial knowledge.

3. Structural Communication

This is a very powerful and flexible method of fixed-response assessment which can range in use from the checking of facts and simple relationships to the construction of "objectively markable essays".

The earliest ideas for this kind of assessment are found in the work of Egan [7] and have since been developed and expanded by other workers, particularly in my own research group. The name, Structural Communication, which Egan used, aptly describes how the method works. The students are presented with a random array of information and are asked to rearrange the array in such a way as to communicate their knowledge and understanding to the examiner. The structure which the students impose on the given information reveals their ideas and their reasoning. The array of information can be presented as a grid of numbered boxes each containing a piece of information, or the information may appear as a series of numbered statements presented one below the other. By a series of examples, I should like to show the range of possible uses of the technique.

(a) Categorising

1 ATLANTIC	2 JAPAN	3 MEDITERRANEAN
4 ARCTIC	5 RED	6 INDIAN
7 PACIFIC	8 TASMAN	9 ANTARCTIC

Figure 4.6

The grid contains the names of oceans and seas. Answer the following questions by selecting appropriate boxes and writing their numbers in the spaces provided.

Note that any box may be used several times to answer different questions.

Question 1. Which of these are in **both** the northern and southern hemispheres?

Question 2. Which of these are around the coast of Canada?

Question 3. Which of these are around the coast of Israel?

Question 4. Which of these are in the southern hemisphere only?

This is testing knowledge and relationships. There is no indication of only one correct response (as in multiple-choice), because one or more box numbers may be required to answer each question and so guessing is much reduced. The same box can be used several times as part of the answer to a number of questions and so answering by elimination is avoided. Partial knowledge is allowed for. However, there is one

drawback which must be countered. If students are given credit for their correct choices and go unpenalised for wrong choices, the clever ones could give all nine boxes as the answer to all the questions. Egan suggested a correction factor to get round this. Suppose that the correct answer to Question 1 above was Atlantic, Pacific and Indian (i.e. boxes 1, 7, 6). There are three "correct" boxes and six "incorrect" boxes. His scoring system was:

$$\text{Score} = \frac{\text{Number of correct boxes chosen}}{\text{Number of correct boxes available}} - \frac{\text{Number of incorrect boxes chosen}}{\text{Number of incorrect boxes available}}$$

A student who responded to Q.1 above with 1, 6, 7 would have:

$$\text{Score} = \frac{3}{3} - \frac{0}{6} = 1.0$$

A student who responded 1 and 7 and omitted 6 would have:

$$\text{Score} = \frac{2}{3} - \frac{0}{6} = 0.7$$ (Partial knowledge is rewarded).

However, if a student's response was 1, 2, 7 the score would be given by:

$$\text{Score} = \frac{2}{3} - \frac{1}{6} = 0.5$$

The student, being clever, who chose all the boxes would have a score of

$$\text{Score} = \frac{3}{3} - \frac{6}{6} = 0$$

This arithmetical procedure is a little tedious to do by hand, but computers mark this effortlessly. Students, and teachers, are not too happy with fractional scores, and so all of these can be multiplied by some simple factor such as ten to give whole numbers. And so the first student above would score 10 x 1 = 10; the second would get 7; the third get 5 and the last still get 0.

Programs are commercially available which remove any chore and apply any weighting the examiner desires [8]. A full treatment for scoring Structural Communication questions is given in the Appendix to this section.

(b) Pattern seeking

This is an extension of categorisation in which the examiner gives examples and non-examples of some pattern and asks the student to deduce the pattern and seek for other examples.

1 Nile	2 Mons Graupius	3 Hastings
4 Stamford Bridge	5 Waterloo	6 Sherrifmuir
7 Trafalgar	8 Bannockburn	9 Moscow

Figure 4.7

In these questions about battles you will be given two examples and one non-example of an idea the examiner has in mind. You are asked to work out this idea and show that you have it by choosing further examples from the grid.

Question 1. Examples are in boxes 1 and 5, but box 3 is a non-example. Choose any other examples from the grid.

The reasoning should be, Nile and Waterloo are examples of the Napoleonic War, but the Battle of Hastings is not Napoleonic. Trafalgar and Moscow were also battles in the Napoleonic Wars and so choose boxes 7 and 9.

Question 2. Boxes 2 and 6 are examples; box 4 is a non-example. Choose any other examples from the grid.

The reasoning might be Mons Graupius and Sherrifmuir are battles which took place in Britain, but Stamford Bridge (the non-example) is also in Britain. However, examples 2 and 6 were battles in Scotland and so the other example is Bannockburn, box 8.

(c) Sequencing

In this version, the student is asked to choose relevant boxes and then to sequence the responses to communicate more fully.

1 TIN	2 IRON	3 COPPER
4 BRASS	5 MAGNESIUM	6 LEAD
7 COBALT	8 BRONZE	9 SODIUM

Figure 4.8

Question 1. Select the metals from the grid which are TRANSITION metals and arrange them in order of increasing atomic number.

Question 2. Which metals are in the alloy in box 8? List them with the one of highest proportion first.

Question 3. Arrange the elements in boxes 2, 3 and 9 in order of their date of discovery with the oldest first.

The scoring for this type of question is more complex because it must have two parts, the choice of boxes and the sequence order. The choice can be scored as explained before, but the sequence presents a problem. There are a number of methods available for computer marking.

If the correct choices have been made (no more and no fewer), the computer compares the sequence with one provided by the examiner. A perfect fit gets full marks, a complete reversal gets no marks. If two adjacent responses are in the correct order, but the remainder are out of order, a partial mark is given.

However, if the original choices were incorrect (for example, by the inclusion of a wrong choice) this is penalised by the first stage of the scoring and then ignored in the sequencing. This is tedious to do by hand, but is easily achieved by machine marking.

So far all the examples have been presented as one word in each box, but the boxes can contain diagrams, pictures, sentences, formulae (mathematical or chemical) or structures. This increases their flexibility and usefulness (Figure 4.9).

1 CH_3OH	2 $CH_3C(OCl)=O$	3 $CH_3C(H)=O$
4 C_2H_5Cl	5 $CH_3CH(Cl)C(=O)OH$	6 C_2H_5OH
7 $CH_3C(=O)OH$	8 $C_2H_5C(=O)H$	9 $CH_3C(OCH_3)=O$

Figure 4.9

Question 1. Select the box(es) which contain alcohols.

Question 2. When the substance in box 6 is oxidised it can produce more than one product. Which boxes contain these products?

Question 3. If this oxidation is carried out in stages, arrange your choice of products in order of their occurrence.

Question 4. Select the box(es) which contain acid chlorides.

Question 5. Which boxes contain the substances which are used to make the compound in box 9?

(d) The "Objective Essay"

Conventional extended answer questions are good for assessing reasoning and the ability to marshal material into a logical whole. They are, however, difficult to mark consistently. Something approximating to an extended answer question can be achieved by Structural Communication techniques. Indeed Structural Communication is at its best when used in this way. The setting of such questions is most easily done in this way.

Ask yourself a question which would need three or four sentences (or ideas) strung together to answer it. Reduce your answer to these basic ideas and insert them randomly into the blank grid. Now ask yourself a second and related question and proceed as before. Preferably some of the ideas needed to answer this second question were also necessary to answer the first.

Finally, ask yourself a third question related to the first two and complete the grid.

You may need to extend the grid beyond the nine boxes we have used in the previous examples, but twelve, or at most sixteen, boxes can be used. An example might look like this, with question fragments represented by numbers in the boxes (Figure 4.10).

A This idea is needed to answer questions **1 and 3**	B This idea is needed to answer question **2 only**	C This idea is needed to answer question **3 only**
D This idea is needed to answer question **2 only**	E This idea is needed to answer question **3 only**	F This idea is needed to answer question **1 only**
G This idea is needed to answer questions **1 and 2**	H This idea is needed to answer questions **1 and 3**	I This idea is needed to answer questions **2 and 3**
J This idea is needed to answer question **3 only**	K This idea is needed to answer question **2 only**	L This idea is needed to answer question **1 only**

Figure 4.10

The students are now given your three questions and the grid filled in with your answer fragments. They are instructed to answer each question by selecting the necessary fragments and presenting them by arranging the box letters in a logical order. The odds against guessing in this situation are very large and can be discounted.

At this level of sophistication, there may be more than one logical order permissible, but the computer (or even the hand marker) can cope with this. The scoring of the selection is easily done as before, but the scoring of the sequencing is more complex, but not too difficult.

This is nearly a machine-markable essay, testing two of the key skills which a normal essay requires; selecting relevant material from irrelevant and presentation in a logical order. Students do not find this an easy option to the conventional essay, but the marker's burden is considerably eased!

A form of Structural Communication, which is now available commercially [8], presents the fragments as a succession of statements down the screen, with the question at the top. The student is asked to study each fragment and decide whether or not it is needed to answer the question. The student selects the fragments and all the others disappear from the screen for the time being. The full screen can be recalled if there is any doubt. The student then "drags and drops" the fragments around the screen to obtain the logical sequence required and, when satisfied, declares the question answered. This is then repeated for each of the questions. Optically this is probably better than the grid, in that the student's final decisions are made without the intrusion of the irrelevant pieces.

We have now looked at three techniques which try to avoid the main drawbacks of fixed-response testing; (i) guessing (ii) lack of information about reasoning and (iii) no allowance for partial knowledge. There is, however, one problem which has not been overcome and that is the need for freedom of expression by students who are more intellectually mature. They want to be able to show their own ideas and their own reasoning and to have room for original insights. The remedy which makes assessment open and "congenial" to all students, is to use a mixture of assessment tools. Both open-ended and fixed-response testing have a place with the proportions changing in favour of the open-ended as the students mature. In fact, since students are driven by assessment, this change in the blend towards more open-ended testing may be one of the tools necessary to encourage the maturation.

Much effort is being expended on curricular innovation without the same effort being applied to assessment innovation. It is worrying to see academics, being driven by various forces to "modernise" their teaching, who then confuse fixed-response with "modern" assessment and effectively neutralise any good they may be achieving in the curriculum.

We have now explored assessment methods in terms of producing individual questions, but we must now move on to the process of **test design** in which questions become papers: measurement tools for our students.

Appendix: Scoring a Structural Communication Question

As mentioned earlier in the section, this is done in two stages: a score for the selection of pieces of information to answer the question and another score for any sequencing which is required.

This can best be shown by an example.

Suppose we have a nine box grid and that the answer to a question is given by the boxes 2, 6, 7 and 9 and the most logical sequence is 6, 7, 2, 9.

Now let us score several student attempts.

Student A has chosen boxes 2, 7, 8 and 9 and has sequenced them as 8, 7, 2, 9.
Choice Score = 3/4 - 1/5 = 0.75 - 0.2 = 0.55

Since there is the possibility of a negative score, this can be eliminated by adding 1 to the choice score before multiplying it by some factor (such as 5) to produce a number which will be recognisable to the student.

Final choice score = (0.55 + 1) x 5
= 1.55 x 5 = 7.75
or rounded up to 8.

The perfect score would have been 10.

Sequence score is based upon a set of yes/no decisions. In this case the most logical order is 6, 7, 2, 9. The questions are -
Does 6 come before 7 (Y/N) and are they adjacent (Y/N)
Does 7 come before 2 (Y/N) and are they adjacent (Y/N)
Does 2 come before 9 (Y/N) and are they adjacent (Y/N)

The student has chosen 8, 7, 2, 9. Now let us apply the test.
Does 6 come before 7 (N) and are they adjacent (N)
Does 7 come before 2 (Y) and are they adjacent (Y)
Does 2 come before 9 (Y) and are they adjacent (Y)

This would score one point for each Y and so would give 4 marks out of a possible 6. The penalty here is because 6 was omitted in the original choice.

The total for the question out of a possible 16 marks is the choice score + sequence score.
= 8 + 4 = <u>12</u>

Student B Chose 2, 6 and 9 only and sequenced them as 2, 9 and 6.
Choice score = 3/4 - 0/5 = 0.75
Adjusted score = (0.75 + 1) x 5 = 1.75 x 5
= 8.75 = 9.0
Sequence score :
Does 6 come before 7 (N) and are they adjacent (N)
Does 7 come before 2 (N) and are they adjacent (N)
Does 2 come before 9 (Y) and are they adjacent (Y)
This would give a sequence score of 2.
Total score = 9 + 2 = 11 out of a possible 16 marks.
This student fared badly on the sequencing and the omission of 7 caused trouble.

Student C chose 3, 4, 5, 6 and 7 and sequenced them as 6, 7, 3, 4, 5.
Choice score = 2/4 - 3/5
= 0.5 - 0.6 = -0.1
= (- 0.1 + 1) x 5
Adjusted = 0.9 x 5 = 4.5

Sequence score
Does 6 came before 7 (Y) and are they adjacent (Y)
Does 7 come before 2 (N) and are they adjacent (N)
Does 2 come before 9 (N) and are they adjacent (N)
The score here is 2.
The total score for the question is 4.5 + 2 = 6.5 (out of 16)

The weighting can be changed in either the choice score or sequence score to achieve the balance the teacher wants. It is wise to inform the students of this balance in advance so that they can share the teacher's view of the relative importance of choice and logical presentation.

SECTION 5: DESIGNING A TEST

Test Specification

For a fixed-response paper the questions to begin with are new but, as time goes by, a bank of questions, which have been pretested, can be laid down. Questions which have functioned well can be banked along with their "statistics" such as facility value (the fraction of the class who got it correct the last time), discriminating power (ability to separate students), response pattern (which distracters worked), topic area, Bloom category and question format.

From this rich information it is possible to specify a paper in such a way as to predict its outcomes in terms of mean score and discrimination leading to a predictable distribution. The test compiler can also ensure proportional cover of topics and skills to meet any specification and to make the paper reflect the time spent on each topic. To do this, without repeating the same paper from year to year, requires a very substantial bank.

To illustrate this procedure, let us specify a twenty question paper of fixed-response questions. In practice such papers are much longer, but this reduced exercise will enable you to experience the process.

Specification

This is just a specimen for this exercise and does not imply any "ideal" specification.

1. 20 questions all worth one mark.

2. Average facility value = 0.55
   (i.e. average mark on test as a whole should be close to 55%).

3. Acceptable limits for facility values 0.3 to 0.8
   (i.e. not too difficult and not too easy).

4. Discrimination factor not less than 0.3
   (i.e. the paper is planned to give a clear order of merit: norm-referenced).

5. The Bloom skills: Recall (recognition), comprehension, application, others in the ratio 4 : 3 : 2 : 1.

6. Each distracter should attract at least 5% of student responses.

7. The relative number of questions should reflect the time spent on each section of the course.

   Time devoted to each section:

   A 8 weeks
   B 10 weeks
   C 2 weeks
   D 4 weeks
   E 6 weeks
   F 10 weeks

   Total 40 weeks

First thoughts

We have only twenty questions to form the test paper. To keep to the specification, 8 of these will have to be recall (recognition), 6 comprehension, 4 application and 2 others.

Also to fit the time requirements 4 will be on topic A; 5 on topic B; 1 on topic C; 2 on topic D; 3 on topic E and 5 on topic F.

Bloom Categories	**Topic**					
	A(4)	**B(5)**	**C(1)**	**D(2)**	**E(3)**	**F(5)**
Recall (8)						
Comp. (6)						
App. (4)						
Others (2)						

Table 5.1

Preparation of a simple table (Table 5.1) will help to keep the picture clear.

Now we must go to the bank (Table 5.2) and look for questions to fit the rest of the specification. This is a fairly typical printout from a question bank.

Statistics for Banked Questions

Course Section	Bank No.	Skill	Facility Value	Discrimination	(a)	(b)	(c)	(d)
A	**1**	**Comp**	**0.7**	**0.5**	**7**	**15**	**71**	**7**
	2	**Recall**	**0.6**	**0.5**	**34**	**5**	**1**	**60**
	3	**Comp**	**0.7**	**0.6**	**69**	**9**	**10**	**12**
	4	**Comp**	**0.5**	**0.1**	**20**	**16**	**45**	**19**
	5	**Recall**	**0.8**	**0.3**	**13**	**78**	**5**	**4**
	6	**App**	**0.5**	**0.4**	**5**	**5**	**53**	**37**
B	**7**	**Recall**	**0.8**	**0.4**	**10**	**4**	**10**	**76**
	8	**App**	**0.6**	**0.4**	**7**	**64**	**7**	**22**
	9	**Comp**	**0.9**	**0.5**	**6**	**89**	**3**	**2**
	10	**Comp**	**0.5**	**0.4**	**5**	**14**	**53**	**28**
	11	**App**	**0.5**	**0.3**	**5**	**7**	**42**	**46**
	12	**Others**	**0.2**	**0.1**	**23**	**37**	**24**	**16**
	13	**Others**	**0.5**	**0.3**	**48**	**12**	**16**	**24**
	14	**Comp**	**0.3**	**0.2**	**40**	**13**	**17**	**30**
C	**15**	**App**	**0.6**	**0.4**	**13**	**8**	**24**	**55**
	16	**Comp**	**0.4**	**0.3**	**17**	**40**	**27**	**16**
	17	**Others**	**0.2**	**0.3**	**20**	**25**	**18**	**37**
	18	**Recall**	**0.8**	**0.4**	**84**	**8**	**3**	**5**
	19	**Recall**	**0.2**	**0.4**	**41**	**24**	**24**	**11**
D	**20**	**Comp**	**0.3**	**0.5**	**24**	**28**	**32**	**16**
	21	**Comp**	**0.4**	**0.4**	**7**	**36**	**17**	**40**
	22	**Others**	**0.4**	**0.5**	**42**	**13**	**27**	**18**
	23	**App**	**0.2**	**0.1**	**16**	**22**	**32**	**30**
	24	**App**	**0.8**	**0.3**	**82**	**3**	**12**	**3**
	25	**App**	**0.6**	**0.4**	**14**	**17**	**63**	**6**
E	**26**	**App**	**0.6**	**0.5**	**27**	**10**	**56**	**7**
	27	**Recall**	**0.4**	**0.6**	**42**	**21**	**18**	**19**
	28	**Comp**	**0.6**	**0.5**	**12**	**61**	**14**	**13**
	29	**Comp**	**0.7**	**0.4**	**10**	**9**	**10**	**71**
	30	**App**	**0.2**	**0.3**	**18**	**29**	**26**	**27**
	31	**Recall**	**0.7**	**0.5**	**20**	**68**	**7**	**5**
F	**32**	**Comp**	**0.3**	**0.4**	**29**	**17**	**24**	**30**
	33	**Others**	**0.6**	**0.3**	**12**	**57**	**22**	**9**
	34	**App**	**0.5**	**0.3**	**52**	**16**	**30**	**2**
	35	**App**	**0.4**	**0.4**	**44**	**10**	**6**	**40**
	36	**App**	**0.1**	**0.5**	**3**	**50**	**12**	**35**
	37	**Comp**	**0.5**	**0.3**	**20**	**47**	**24**	**9**
	38	**Recall**	**0.3**	**0.4**	**33**	**18**	**21**	**28**
	39	**Recall**	**0.4**	**0.4**	**18**	**12**	**35**	**35**
	40	**Recall**	**0.5**	**0.3**	**38**	**48**	**6**	**8**

The underlined figure in the response pattern is the KEY (i.e. correct answer). The value of the KEY is a percentage, which rounded up (or down) should correspond to the facility value. For example, in Q.19, the key is 24%, which rounds down to give a facility value of 0.2 . In Q.13 the key is 48% which rounds up to give a facility value of 0.5.

Table 5. 2

Now looking through the bank there are some questions we have to eliminate because they do not fit the criteria regarding the facility value range (0.3 to 0.8), the discrimination (at least 0.3) and the functioning of the distracters (at least 5%).

Question 2 has a very weak distracter, attracting only 1% of the responses.

Question 4 does not discriminate well enough (0.1).

Question 7 just fails on one distracter which is attracting only 4% of the responses.

Continuing this process we could eliminate Questions 9, 12, 14, 17, 19, 23, 30, 34, 36.

Already twelve of our forty banked questions do not fit the criteria. To approximate to the specification, we may have to relax some of the criteria to allow the test to be constructed to reflect the time and skills distributions.

Using Table 5.1 we can now try to fit the non-eliminated questions into it.

Topic A needs four questions, and only four have survived the elimination process. We must use all four: questions 1, 3, 5 and 6.

Topic B needs five questions, but only 8, 10, 11 and 13 have survived. However, question 7 only narrowly failed and so we will include it to give the five questions needed for the specification, 7, 8, 10, 11 and 13.

Topic C needs one question and we have a choice of 15, 16 or 19.

Topic D needs two questions and we have a choice of 20, 21, 22 and 25.

Topic E needs three questions and we have a choice of 26, 27, 28, 29 and 31.

Topic F needs five questions and the choice is 32, 33, 35, 37, 38, 39, 40.

It now remains to fit the available questions to the skills and to have an average facility value of somewhere near 0.55.

Table 5.3 shows one possible solution to fit the specification.

	Topic						
Bloom Categories	**A(4)**	**B(5)**	**C(1)**	**D(2)**	**E(3)**	**F(5)**	**Totals**
Recall(8)	**Q5 (0.8)**	**Q7(0.8)**	**Q19(0.2)**		**Q27(0.4)** **Q31(0.7)**	**Q38(0.3)** **Q39(0.4)** **Q40(0.5)**	**8**
Comp(6)	**Q1(0.7)** **Q3(0.7)**	**Q10(0.5)**		**Q21(0.4)**	**Q29(0.7)**	**Q37(0.5)**	**6**
App(4)	**Q6(0.5)**	**Q8(0.6)** **Q11(0.5)**		**Q25(0.4)**			**4**
Others (2)		**Q13(0.5)**				**Q33(0.6)**	**2**
Totals	**4**	**5**	**1**	**2**	**3**	**5**	**20**

Table 5.3

Each question is shown with its facility value. The average of these values is 0.53 which is close to the specified 0.55.

Although the bank contained twice as many questions as the test we were trying to construct, it was not easy to meet the specification and it would be very difficult to construct another twenty question paper from the same bank without a large overlap.

It is recommended that a bank should contain at least **ten times** as many questions as will be required for a given size of test. Even then, this would be insufficient if the bank did not contain a wide spread of questions over all the topics and embrace the full range of skills.

If the bank is kept on a computer and the criteria are also built-in, it is possible to assemble a test quickly and randomly, keeping overlap between tests to a minimum.

A sobering thought

All of the material which we have dealt with in the previous sections has to be treated with caution. The processes described tend to indicate that assessment can be made an "exact science" with numbers, statistics and computers. Many people are beguiled by numbers which are more appealing than estimates and guesses about student performance. Indeed, the whole intention of these sections has been to steer the reader towards a more rigorous approach to assessment and to sharpen up a blunt instrument. However, the most sophisticated assessment procedures are entirely useless if the questions themselves are faulty in any way. No amount of statistical manipulation will be able to compensate for questions of doubtful quality. Shredding and reshredding, in the light of the results of questions being tried on students, is the only safeguard we have against self-delusion and unfairness to our students.

In the next section we shall look at the procedures necessary to make sense of results from tests provided that the questions are sound and that the papers have been carefully constructed.

SECTION 6: BEWARE OF NUMBERS

As scientists we are all familiar with numbers, but have you stopped to consider the nature of numbers? Things are seldom what they seem, and numbers are no exception.

There are numbers which are *absolute*. A tape measure will give you numbers for the length and width of your desk. Your bathroom scales will give you numbers for your weight and a map will give you numbers for the height of a mountain. These numbers are all made up of fixed units; metres for length, kilograms (or pounds) for your weight and metres for the height of the mountain. They all have a zero; a point from which the measurement is made. In the case of the tape, it yields a number of centimetres counted from the end of the tape and the marks on the tape have been calibrated against some standard, probably the metre. The bathroom scales have also been calibrated against some standard kilogram. The height of the mountain is taken against a zero, an agreed value for sea level, and the instruments which made the measurement have been calibrated against a standard metre. All of this is elementary common sense, but not all numbers and measurements fit this pattern.

If you have a bowl of mixed fruit; apples, oranges, pears, plums and grapes, you can count all the pieces of fruit, but the total is fairly meaningless. The total number gives no information about the number of apples or grapes. Someone else could have a bowl of fruit with the same total number of pieces as before, but there is no way that we can assume that the composition of the two bowls is the same. The second bowl may have no apples, but extra grapes. However, this total number of pieces of fruit has a zero against which to count; an empty bowl.

A similar, and more meaningless number could be obtained by adding up the wheels in your car. You could add the number of road wheels to the steering wheel, to all the wheels associated with the engine and arrive at a total which would tell you precisely nothing.

What has all this to do with assessment? Be patient for a moment; all will be made plain!

In an examination we award marks for various responses. Can we be sure that the value of any one mark is the same as the value of any other mark given in the test? A mark for an easy bit of recall is not necessarily equivalent to a mark for a piece of "unseen" reasoning. A mark cannot be referred to a fixed standard in the way that a metre on a tape measure can be compared to a standard metre. Test scores are similar to the number of pieces of fruit in a bowl. We add things of different size and character and come up with a composite number which cannot be endowed with too much meaning. There is an added hazard attached to test scores in that they do not have a fixed zero. Of course, a student can score zero on a test, but that does not mean that he knows nothing about the course? It may mean that he knows nothing about what you have asked, but unless your test is totally comprehensive, you have no absolute zero.

There is also the other snag which we met with the bowls of fruit. If on a test two students obtain the same score, there is only the slightest chance that these two

students are exactly (or even fairly) equivalent. One may have gained their total mainly from recall questions while another may have made up their total from other skills.

Let us do some simple arithmetic to confirm this. Suppose we have a test with a total possible score of 10 marks obtained from 10 questions each valued at 1 mark.

There is only *one* way to get 0 marks (fail all the questions) and only *one* way to get 10 marks (pass all the questions).

There are *ten* ways of getting 1 mark (pass one question only) and *ten* ways of getting 9 marks (fail one question only).

There are *forty five* ways of getting 2 marks and *forty five* ways of getting 8 marks. It is getting alarming, but just wait! (Figure 6.1).

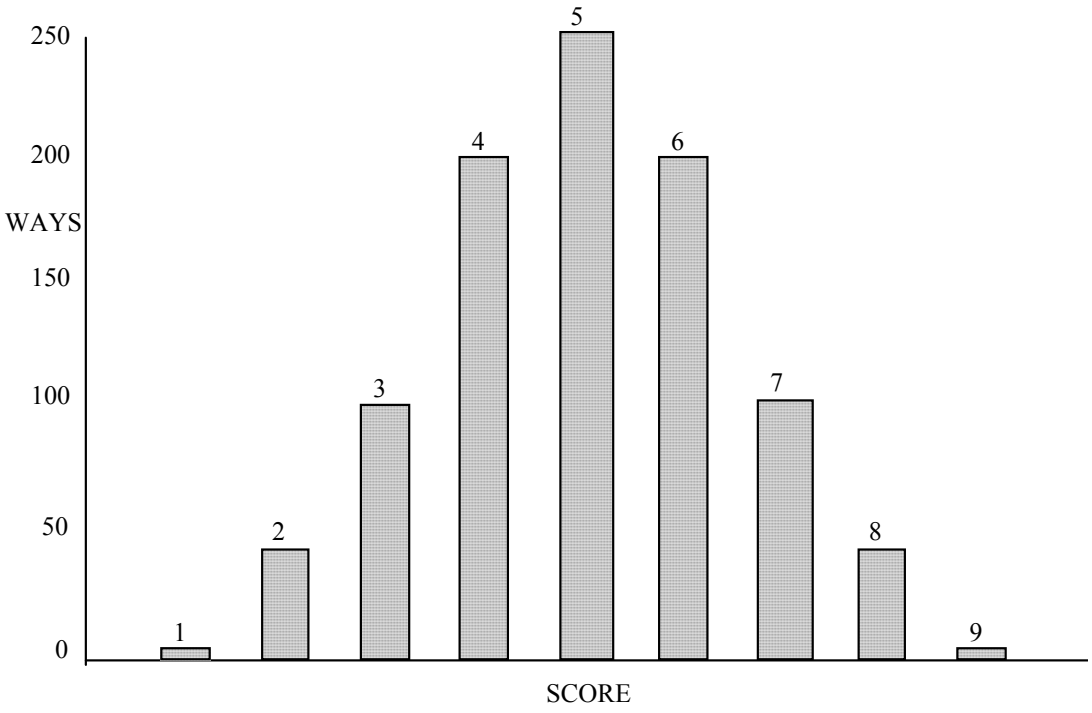


Figure 6.1

What about half marks, 5 out of 10? The staggering fact is that there are 252 ways of assembling the score 5 out of 10. Can you be sure that all of these are equivalent? The best we can do is to acknowledge the limitations on the numbers we obtain from tests and treat them with the greatest caution. As the size of the test increases, this problem gets astronomically worse and a score of 50 out of 100 can be obtained in 1 x 10^{29} ways, that is about a million times Avogadro's Number!! There is no possibility that any two of these are identical although many may be similar. Do not despair, but do be realistic in your use of examination scores. They are numbers of a special kind, not to be confused with absolute measurements of length or weight.

There is a simple device to enable a teacher to detect any student who is gaining total scores in an unusual way and this will be explained later in this section. This will go some way to help us to use test scores intelligently.

Let us sum up so far

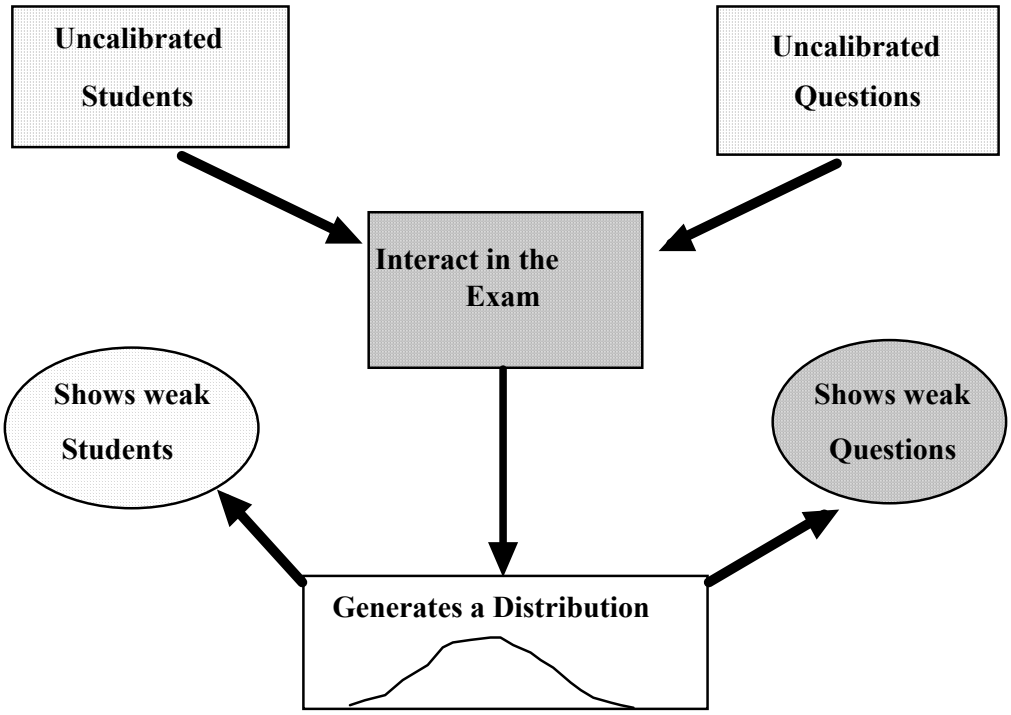


Figure 6.2

A test is meant to be an instrument for measuring the attainment of our students, but as we have shown above, the tests can be uncalibrated tools incapable of making absolute measurements. Yet we are using tests to place a value, a measurement, on the efforts of our students. This is like taking a piece of wood and placing markings on it spaced out unevenly and using it to measure the heights of our students. We will be able to put them in order of height, but the values (the numbers) we attribute to that height will be, at best, unreliable in any absolute sense.

A test, if norm-referenced, will put students in order of merit and will generate some sort of distribution, with few very poor students and few outstandingly good and most of the others bunched in the middle. Do not be surprised if the distribution you obtain is not the expected bell-shaped curve. It might often resemble a two or even three humped camel!

Putting test scores together

Let us try to penetrate this arithmetical jungle a little further before we start to consider how to offset the worst of the problems and handle scores sympathetically and sensibly.

The problems associated with a single test are compounded when we try to combine them with scores from another test. Because a mark does not necessarily equal another mark in the same test, there is even less likelihood of a mark in one test being equivalent to a mark in another. This can be taken further when we consider total marks in two tests. Does it follow that a total of 50 marks in one test is likely to have the same meaning as 50 marks in another test? Almost certainly this will not be so. Most teachers are aware that total scores on projects tend to be much higher than those in written tests. Here the differences are starkly obvious, but even between two written papers there are significant differences.

Let us consider a set of test results very similar to a real set often encountered in my experience. The examination consists of three papers each in a different part of the discipline. Figures 6.3 (a,b,c) shows the distribution curves for the three exams.

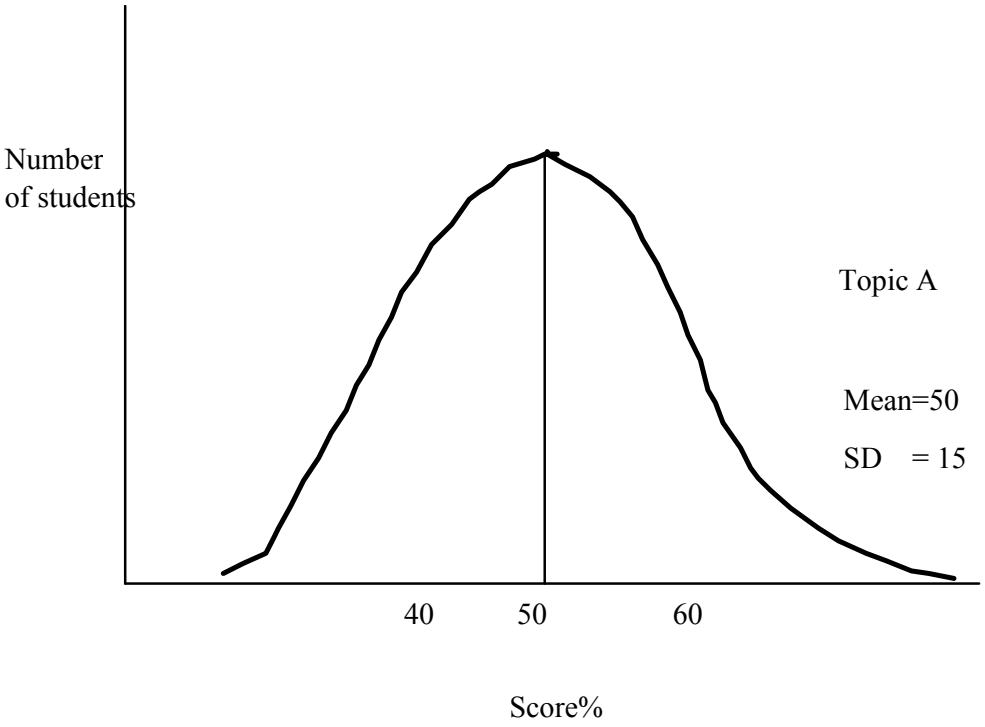


Figure 6.3 (a)

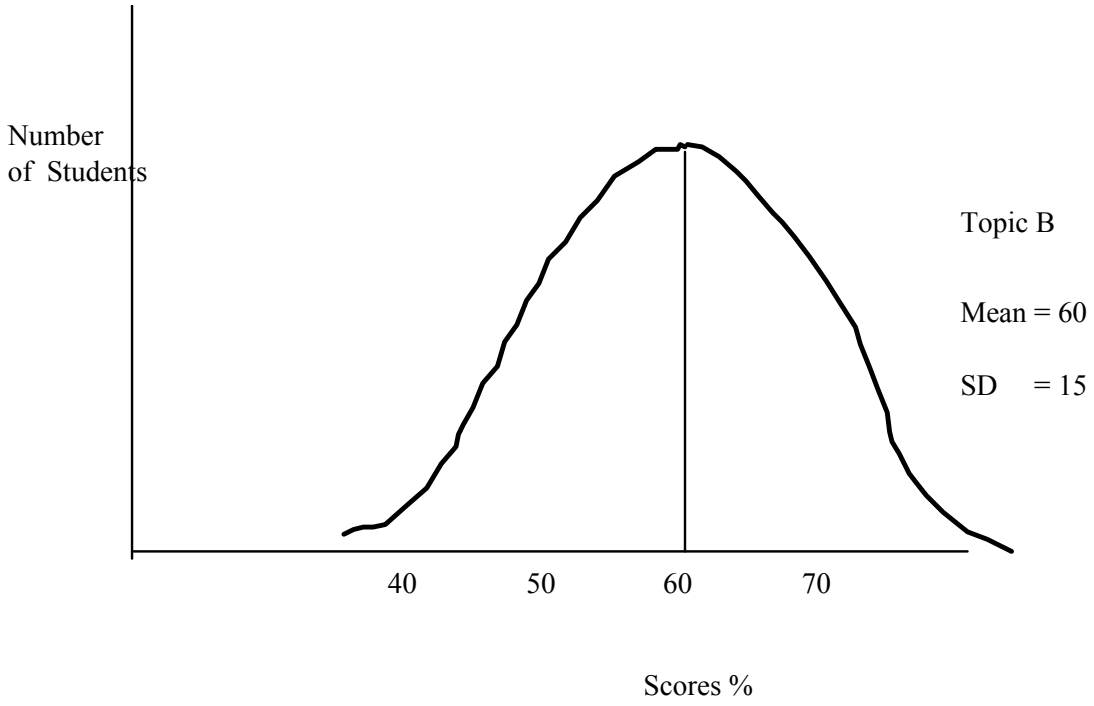


Figure 6.3 (b)

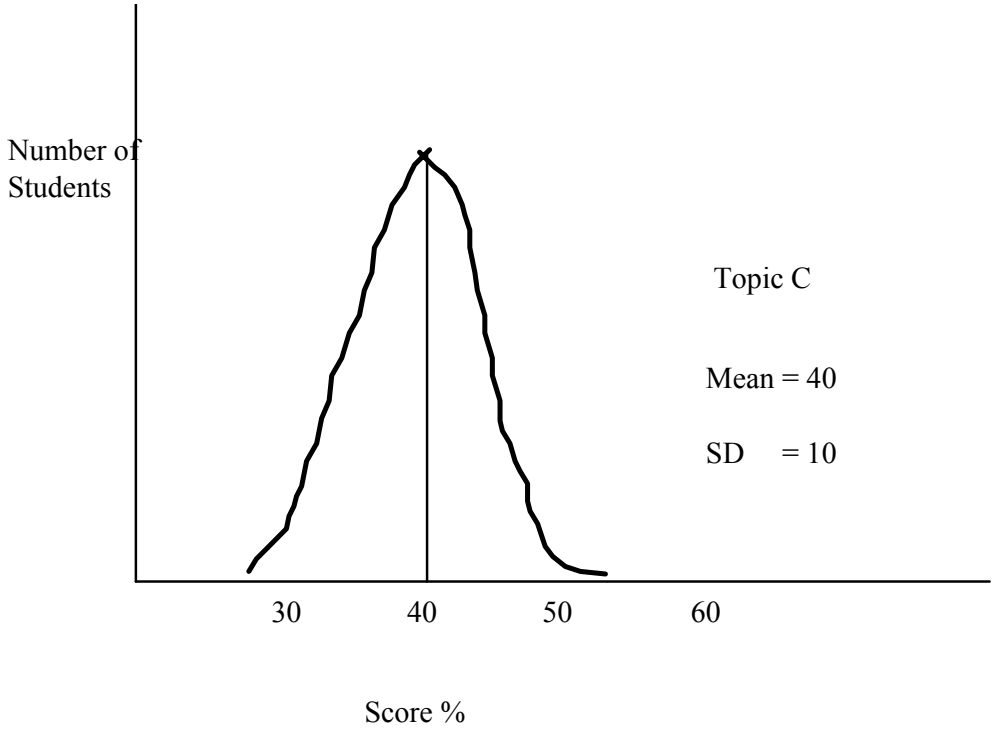


Figure 6.3 (c)

In topic A a score of 50 is exactly on the average, while in topic B a score of 50 is well below average, and in topic C a score of 50 is close to genius level! Clearly a score of 50 does not have the same meaning in all three tests. If, for an individual student, we simply add the scores on the three tests, we are likely to create some strange anomalies. In Table 6.1 we have the raw scores for three students totalled to give a value of 150.

	Topic A	**Topic B**	**Topic C**	**TOTAL**
Student 1	**50**	**50**	**50**	**150**
Student 2	**40**	**50**	**60**	**150**
Student 3	**50**	**60**	**40**	**150**

Table 6.1

Student 3 has obtained the average score on each test and so we might class him as "the average student". **Student 1** has an average performance in topic A, below average in topic B and above average in topic C. **Student 2** is below average in topic A, below average in topic B, but well above average in topic C. These three students are not equivalent although their total scores are the same.

If only we could make each test have the same average, and adjust the shape of the curves to superimpose neatly on each other, we might be able to take account of the differences between the students. In Figures 6.3 there is a number given for SD, the Standard Deviation. This is a measure of the spread of the scores. The bells for topics A and B are similar, but B is over to the right. This is shown by both having the same SD = 15 but the mean (average) for B is 10% higher. The bell for topic C is much narrower than those for A and B and is displaced to the left of them. This is expressed in a smaller SD = 10, and a lower mean.

It is arithmetically easy to recalculate the values for the three curves to allow them all to have the same average and the same distribution shape (same SD). This would then allow us more legitimately to add the scores for individuals and take account of the differences in their performance. If you want to do this, ask your computer to calculate Standard Scores for you. If you do not want to know about the calculation go directly to Table 6.2.

For the reader who wants to know what is going on arithmetically, let us calculate the standard scores for our three students.

Let us decide to use the mean and SD for the topic A curve as the "standard"; but we could use any of them as standard. To calculate the Standard Score we use the following formula.

$$\text{Standard Score} = \left[\frac{\text{Raw score - Raw mean}}{\text{Raw SD}} \text{ x Standard SD} \right] + \text{Standard Mean}$$

We will use this to calculate the new Standard Scores for the students. For topic B, the SD is the same as topic A (the Standard) and so the formula simplifies.

Standard Scores for **topic B** become = Raw score - raw mean + new mean

Student 1 becomes 50 - 60 + 50 = 40

Student 2 becomes 50 - 60 + 50 = 40

Student 3 becomes 60 - 60 + 50 = 50

For topic C, we will have to take account of the change in SD.

Student 1 becomes**:**

$$\left[\frac{50 - 40}{10} \text{ x } 15 \right] + 50 = 65$$

Student 2 becomes**:**

$$\left[\frac{60 - 40}{10} \text{ x } 15 \right] + 50 = 80$$

Student 3 becomes**:**

$$\left[\frac{40 - 40}{10} \text{ x } 15 \right] + 50 = 50$$

Now let us look at the new Standard Scores Table (Table 6.2).

	Topic A	**Topic B**	**Topic C**	**Totals**
Student 1	50	40	65	155
Student 2	40	40	80	160
Student 3	50	50	50	150

Table 6.2

Since all three papers now have an average of 50, Student 3 (the average one) has all three scores of 50 and a total of 150 as before. However, Students 1 and 2 have new totals which reflect their performances more adequately. The students are no longer equal, and the differences may be big enough to award them different final grades. At honours level, such calculations can change a 2.2 into a 2.1 (and vice versa) and even change a 2.1 to 1st class and vice versa.

Another thought about adding raw scores

If you are trying to place students in order of merit, it is important to look at the spread of marks in the tests you propose to add together. An example will illustrate this (Table 6.3).

Students		A	B	C	D
Paper 1	Score	10	21	38	50
	Rank	4th	3rd	2nd	1st
Paper 2	Score	100	95	85	80
	Rank	1st	2nd	3rd	4th
Papers 1 + 2	Score	110	116	123	130
	Rank	4th	3rd	2nd	1st

Table 6.3

In this hypothetical case the order of merit in Paper 1 is exactly reversed in Paper 2, but the combined order of merit is the same as that in Paper 1. Why should this be so?

The spread (range) of marks in Paper 1 goes from 10 to 50, that is a range of 40 marks. In Paper 2 the range is 100 to 80, or 20 marks. The Paper with the larger range has a stronger effect upon the combined order of merit than the paper with the smaller range. This distorts the outcome and denies student A the credit for the top place in Paper 2. Making allowances for the differences in spread would leave an order of merit with student B marginally first, students A and D joint second and C just third. This is probably a truer reflection of the situation.

The discussion we have had about putting scores together and the measures we have suggested are better than adding raw scores, but they do not solve the underlying problem that scores from tests are not amenable to the normal arithmetical procedures which we would apply to ordinary numbers. They are indicators, but not rigid criteria for decision making about a student's progress. They do not replace professional judgment, but are aids to inform such judgment.

The magic of 50%

The blind application of a fixed pass mark of, say, 50% is logically not tenable. It would be by the wildest fluke that 50% on one examination would be comparable with 50% on another. However, one might get somewhere near this if a set of prespecified and pretested questions were used in a fixed-response paper as we indicated in Section 5.

Having said this, 50% still holds unquestioned sway as the pass mark in many institutions. How can uncalibrated tests possibly be a standard unless we believe in some sort of professional infallibility? It only takes an ambiguity in the wording of one question to depress the mean score for a test by a significant amount, making the meaning of 50% as a standard, useless. If we are dealing with large numbers (several hundreds) of students coming from similar age and ability cohorts from year to year, their performance is more likely to be a reliable standard then any test can be. The examination boards handling very large numbers (in the thousands), can use the examination to generate a distribution curve and order of merit. The pass mark is then set as that score which allows a fixed proportion of the students to pass. Appeals are then allowed for scores a few percent below that mark and papers may be re-marked to allow justice to be done, and a few more lucky candidates to succeed. This procedure is not possible for smaller samples and professional judgment must be made, taking account of the flaws exposed in the paper and of what we know from other observations about our students. The 50% pass mark is well ingrained in the educational psyche of teachers and students: 51% brings joy while 49% spells disaster. It is hard to see why 50% has gained this special status. It is hardly an indicator of a satisfactory mastery of a course, if a student can amass only half of what was supposed to be learned. This is even worse when choice is allowed in a paper, for example, five questions out of eight. Assuming that the eight questions adequately covered and examined the objectives and outcomes of the course, 50% represents a "competence" in half of five eighths of the course! How can this be a good grounding for proceeding to the next part of the course? How can this be adequate licence for professional practice?

Bad students or a bad exam?

It is very convenient, when a set of disastrous exam results occur, to blame the students. "It is a bad group this year" seems to mollify the teacher and maybe even the administrator, but this cannot always be so. For small groups of students, it is perfectly possible for this to be a reasonable explanation. One or two exceptionally weak or outstandingly good students can make very significant differences to the average score in an examination where the class size is small. When the class size is large, into the hundreds, such swings are less likely, and major changes in averages and distributions are more difficult to explain. The changes may arise from a change in student intake or from the quality of the test or the teaching or a blend of all three. It is possible to do a rough check on the test and on the student sample to help to locate the source of the variation. When a fixed response test is made up of newly minted questions, it is almost impossible to find the source of the problem. However, if it is partially (or wholly) made up of banked questions with their previous statistics, we have some way of finding out if the student standard has changed by looking at changes in facility values. If the facility values in the banked questions are generally lower, we can with some confidence attribute the problem to "lower student ability".

If the facility values are fairly constant, except for those on a particular topic in the course, then the teaching needs to be looked at. Although this strategy of "internal marker questions" is by no means fool-proof, it is a useful indicator and safeguard and well worth using.

Norm or Criterion-Referencing

In Section 1 we encountered the ideas behind norm and criterion-referencing. In the present section we have been looking at situations which have mainly been norm-referenced. We have talked about distributions, orders of merit, bell-shaped curves and "pass marks" of 50%.

Criterion-referencing is meant to be different in that it does not depend upon accumulation of marks to arrive at some acceptable total, but rather on a series of pass/fail decisions made about fixed attainment hurdles. However, criterion-referencing is not immune to the problems we have uncovered. The quality of the questions may be variable and so the "heights" of the hurdles may alter, even though they purport to measure the same objectives. If the tests are of the objective, fixed-response type, they suffer from the additional drawbacks we saw in earlier sections. Once more the message is "not to be beguiled by numbers". Interpretation of criterion-referenced results must be subject to professional judgment and common sense. The numbers are at best indicators to help that judgment.

However, one form of criterion-referencing seeks to measure *mastery* of a piece of learning. The teaching is usually in small, compact, specified modules with a set of stated educational objectives or outcomes. The student is deemed to have passed this module when a score of at least 80% has been achieved in a fixed test. The outcome is a pass or fail: no "degrees of pass" are allowed. Failure for the student means either abandoning the module or recycling through it until the level of mastery is achieved. Although the score of 80% can be achieved by a number of combinations of marks, this number is relatively small and so there is some comparability between students reaching the goal.

This mastery system may lend itself to some aspects of education such as the "drill" in learning grammar, or performing a specific kind of calculation, or of flying an aircraft (not 80%, but 100%), but would be useless in many other areas. One could never assess a poem, or literary criticism or legal argument or interpersonal skills in this way. Assessment has to fit what is being assessed despite the limitations we have discussed so far. What is essential is an informed view of the nature of the numbers arising from any assessment and to treat them with the care and caution they deserve.

Getting more information from test scores

Earlier in this section we noted that total scores on a test can be made up in many ways, raising doubt about the equivalence of students with the same scores. We shall now look at one way of recognising these differences and using them to help our students.

We shall use a simple example to show how this can be done. Supposing we have ten students sitting a ten question test, each question being worth one mark. This is the kind of situation which you could meet in a multiple-choice test. Table 6.4 shows the results of the test.

	QUESTIONS										
Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**TOTAL**
1	1	0	1	1	0	1	0	0	1	1	6
2	1	1	1	1	1	1	0	1	0	1	8
3	1	0	0	0	0	0	1	0	0	1	3
4	1	1	1	1	1	1	1	1	0	1	9
5	1	0	1	1	0	0	0	0	1	0	4
6	1	1	1	1	1	1	1	1	1	1	10
7	1	0	1	0	1	0	1	0	1	0	5
8	1	1	0	1	1	1	1	0	0	1	7
9	1	0	1	0	1	1	0	1	0	0	5
10	1	1	1	1	1	1	0	1	0	1	8
FV	1.0	0.5	0.8	0.7	0.7	0.7	0.5	0.5	0.4	0.7	

Table 6.4

You will notice in the right hand column the total score for each student and you will see that some have the same score. Along the bottom of the table are the Facility Values (FV) for each question. Remember that the FV is the proportion of the class getting that question correct. For question 1, everybody got it correct and so the FV = 1.0. In question 2 only five students out of ten got it correct and so the FV = 0.5 (half the class).

We are now going to do two simple rearrangements of this data. We will arrange each of the horizontal rows of figures so that the one with the highest total comes to the top, the lowest to the bottom and the others arranged in order in between. This will give us the results in Table 6.5.

	QUESTIONS										
Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**TOTAL**
6	1	1	1	1	1	1	1	1	1	1	10
4	1	1	1	1	1	1	1	1	0	1	9
2	1	1	1	1	1	1	0	1	0	1	8
10	1	1	1	1	1	1	0	1	0	1	8
8	1	1	0	1	1	1	1	0	0	1	7
1	1	0	1	1	0	1	0	0	1	1	6
7	1	0	1	0	1	0	1	0	1	0	5
9	1	0	1	0	1	1	0	1	0	0	5
5	1	0	1	1	0	0	0	0	1	0	4
3	1	0	0	0	0	0	1	0	0	1	3
FV	1.0	0.5	0.8	0.7	0.7	0.7	0.5	0.5	0.4	0.7	

Table 6.5

Now comes the second operation. Look at the columns in Table 6.5 and arrange the data in them so that the column with the highest FV (1.0) is at the left and the column with the lowest FV (0.4) is at the right with the others arranged in order in between. Where Facility Values are the same, their order does not matter. This will now rearrange to give Table 6.6.

	QUESTION										
Student	**1**	**3**	**4**	**5**	**6**	**10**	**2**	**7**	**8**	**9**	**TOTAL**
6	1	1	1	1	1	1	1	1	1	1	10
4	1	1	1	1	1	1	1	1	1	0	9
2	1	1	1	1	1	1	1	0	1	0	8
10	1	1	1	1	1	1	1	0	1	0	8
8	1	0	1	1	1	1	1	1	0	0	7
1	1	1	1	0	1	1	0	0	0	1	6
7	1	1	0	1	0	0	0	1	0	1	5
9	1	1	0	1	1	0	0	0	1	0	5
5	1	1	1	0	0	0	0	0	0	1	4
3	1	0	0	0	0	1	0	1	0	0	3
FV	1.0	0.8	0.7	0.7	0.7	0.7	0.5	0.5	0.5	0.4	

Table 6.6

You will notice that the apparently random arrangement of zeros and ones in Table 6.4 have now yielded an interesting pattern in Table 6.6. In general the ones have migrated to the top left of the table and the zeros to the bottom right of the table.

There are exceptions and we shall deal with these in a moment. In fact these exceptions give us most important information.

Despite what was said earlier about the many ways that students can theoretically compile their scores from different questions, the reality is much simpler. There are some questions which all students can do and some which few can do and these few students will tend to have the highest total scores. Ideally one might expect a given student to score well on questions of high FV and fail on those of lower FV. The best student at the top of the table has succeeded in all the questions; the next student has passed all the questions except the last. However, when we come to the next two students with a total of 8, the pattern is not perfect. The further down we go, the less clear is the pattern. In an "idealised" situation all the ones would be in the top and left of the table and all the zeros would be at the bottom and right of the table, but the deviations from this are informative.

First of all, let us look at the **vertical columns** in Table 6.6. In general the zeros, as expected, come at the foot of the columns, but the columns for questions 7 and 9 are anomalous. Zeros are coming near the top and ones near the foot. In question 9, four of the best students have answered incorrectly and three of the weaker students have succeeded. This is an example of a question showing *negative discrimination*. In Section 3 we looked at the idea of discrimination and indicated how it was calculated. For a question to discriminate usefully, the facility value of that question, for the top third of the class on the tests as a whole, should be greater than that for the bottom third.

In the case of question 9, the facility value for the top four students is one in four, 0.25. For the bottom four students, the facility value is two in four, 0.50. The discrimination for question 9 is 0.25 - 0.50 = -0.25. Something about this question has made it easier for weak students than for strong ones. Perhaps none of the options was satisfactory to the strong students and they just guessed, while the more naive, weak students found a satisfactory option. For whatever reason this anomaly has occurred, question 9 needs to be examined by a shredding team for some faults. In the case of question 7, the discrimination is 0.5 - 0.5 = 0. This question has not contributed to creating separation between students, but neither has question 1. However, one might justify question 1 as an easy starter for all, but question 7 is harder to justify. In summary, looking down columns in this table helps us to see which questions have anomalous properties and need to be looked at again.

Now let us scan Table 6.6 **horizontally**. The two students who scored 8, obtained their total in the same way. However, the two students who scored 5 obtained their totals very differently. They show the same behaviour for the first four questions, but the pattern changes from then onwards. Student 7 has obtained two of the five successes from the most difficult questions and has failed in four moderately easy questions. Is this an able student who has missed some work through illness or are the fail questions based on topics for which the student has some misapprehension? The help this student needs is now revealed for the teacher who wants to offer diagnostic support.

Looking at the rows indicates students with particular needs while examining the columns shows questions in need of rewriting.

To handle tables of this kind (Tables 6.4 - 6.6) would be tedious for larger classes, but the procedure can be automated by entering the results into a computer spreadsheet and getting it to rearrange the data into a convenient form.

This section has dealt with numbers and the traps for the unwary. Its intention is not to depress examiners, but to help them to take a realistic view of what they are doing during the testing process. To ignore the warnings about the misuse of uncalibrated assessment instruments is to delude ourselves and to be unfair to our students. The measures suggested in this section may seem to be tedious and unconventional, but they attempt to sharpen, by more than a few degrees, an otherwise blunt instrument to make assessments yield more valid and reliable results on which sound professional judgments can be made.

Final thoughts

Fixed response testing is much more than cobbling together a few multiple-choice questions and churning the students' responses through the computer. If the testing is to have any meaning, there are skills to be acquired by the setter which may seem tedious to apply but are part of being a thoroughly professional examiner. It is not enough to be an expert in a given scientific field and to assume that the teaching and examining skills will arrive by osmosis. To be worthy of the privilege of teaching at an advanced level, we must strive to be professional in all aspects of our responsibility. It is hoped that this little guide will contribute to the realisation of that goal.

REFERENCES

1. B S Bloom (1956) Taxonomy of educational objectives, Handbook 1, Cognitive Domain, London, Longmans

2. J Handy, A H Johnstone (1973) How students reason in objective tests, Education in Chemistry, **10**, 99

3. P Tamir (1990) Justifying the selection of answers in multiple-choice items. International Journal of Science Education **12**, 563-573

4. A Ambusaidi, A H Johnstone (2000) Fixed response - What are we testing? CERAPIE, **1**, 323-328

5. JRT Cassels, A H Johnstone (1984) The effect of language on student performance in multiple-choice tests in chemistry. Journal of Chemical Education, **61**, 613-615

6. S Friel, A H Johnstone (1978) Scoring systems which allow for partial knowledge, Journal of Chemical Education, **55**, 717-719

7. K Egan (1972) Structural Communication: a new contribution to pedagogy, Programmed Learning Educational Technology, 63-78

8. D Mackenzie (1997) TRIAD; a computer based assessment software, Centre for Interactive Assessment Development, University of Derby

Assessment forms an integral part of the teaching and learning process. It is applied during a course (formative assessment) to help students and teachers to take a realistic view of progress and to catch misunderstandings at a stage when they can be easily rectified. It is also applied at the end of a course (summative assessment) to ensure that the student has not only learned the bits, but can also piece the bits together into a coherent whole.

This guide is largely concerned with the form of assessment called Objective or Fixed- Response testing, but this is only part of the range of assessment tools available. It would be a mistake to think that this form of testing is useful or appropriate for the measurement of all the skills we hope to develop in our students. However, objective testing has clear advantages for the assessment of some skills, not least its ease of scoring.

To clarify where these advantages lie, we shall have to consider the assessment process more widely in the early part of the guide.

The bulk of the guide is devoted to the development of the expertise necessary to design questions, to construct papers and to handle the results. The effort expended to gain this expertise will pay off in terms of better and sharper assessment tools which will help our students to take a realistic view of their progress and which will keep us from self-delusion about our teaching.

The examples of questions in the text have been purposely kept at a low content level to illustrate the method under discussion. This should not be taken to imply that fixed response testing is trivial or incapable of being used to test at all stages at tertiary level.

Alex Johnstone was formerly the Head of the Centre for Science Education, University of Glasgow.

LTSN Physical Sciences

. . . supporting learning and teaching in chemistry, physics and astronomy

Department of Chemistry
University of Hull
Hull
HU6 7RX

Phone: 01482 465418/465453
Fax: 01482 465418
Email: ltsn-psc@hull.ac.uk
Web: www.physsci.ltsn.ac.uk