1 Meanings, processes and properties of assessment

This chapter lays the ground for later chapters by setting out the meaning of terms that are used to describe, analyse and evaluate assessment procedures and systems. In particular it makes clear that the word 'assessment' is used to refer to the process of gathering, interpreting and using evidence to make judgments about students' achievements in education. The term 'evaluation' is reserved for this process of using evidence in relation to programmes, procedures, materials or systems. It also makes explicit the meaning of 'assessment by teachers' or 'teachers' assessment', terms that will feature frequently throughout this book.

A framework of variables within seven main components of assessment is offered as a way of describing different ways of conducting assessment. Finally, properties that need to be considered in making decisions about how to conduct summative assessment are proposed: validity, reliability, impact and use of resources.

Introduction

Any discussion of assessment inevitably involves reference to concepts that are special to the subject – jargon to those not familiar with the terms. Whilst trying to avoid this as far as possible, it is important to use words with some precision in developing arguments relating to topics as complex as assessment. So, although it makes a rather dry start to the book, the matter of terminology used in discussing assessment cannot be avoided. Words such as assessment, evaluation, testing, performance, achievement, formative, summative and so on will have to be used in this book and, without intending to suggest what is correct or incorrect usage, it is essential that their meaning as used here is clear and consistent. It is best to get this done sooner rather than later and to be candid about how difficult it is to be precise.

As part of this clarification, the second section of this introductory chapter looks at how types of assessment can be described in terms of the various ways in which it can be carried out. In the third section, we consider some key properties that ought to be taken into account when evaluating the effectiveness and value of any type of assessment. These properties are revisited throughout the book, in the course of providing evidence and arguments for proposing reform in assessment systems that rely heavily on external tests and examinations.

Meanings

Assessment and Evaluation

Assessment and evaluation both describe a process of collecting and interpreting evidence for some purpose. They both involve decisions about what evidence to use, the collection of that evidence in a systematic and planned way, the interpretation of the evidence to produce a judgment, and the communication and use of that judgment. The evidence, of whatever kind, is only ever an indication or sample of a wider range that could be used.

The terms 'evaluation' and 'assessment' in education are sometimes used with different meanings, but also interchangeably. In some countries, including the USA, the term 'evaluation' is often used to refer to individual student achievement, which in other countries including the UK is described as 'assessment'. In the UK 'evaluation' is more often used to denote the process of collecting evidence and making judgments about programmes, systems, materials, procedures and processes; 'assessment' refers to the process of collecting evidence and making judgments relating to outcomes, such as students' achievement of particular goals of learning or teachers' and others' understanding. The processes of assessment and evaluation are similar but the kinds of evidence, the purpose and the basis on which judgments are made, differ. Our concern here is assessment in this latter sense; that is, the evidence is about what students can do, what they know or how they behave and the judgments are about their achievements.

Assessment Systems

'System' implies a whole comprised of parts that are connected to each other. In the case of assessment the system will include: procedures for collecting evidence, its use for different purposes, and how it will be used for individual reporting, certification and selection; system monitoring at local and national levels; the use of measures of performance of students in the accountability of teachers and schools; the role of teachers in assessment, both formative and summative; the moderation of teachers' judgments;

and the way in which evidence from different sources, such as assessment by teachers and external tests, is combined.

We need look no further than either side of the border between Scotland and England to appreciate the difference between assessment systems. In Scotland summative assessment of pupils up to the age of 15 or 16 is for internal school use only. There is a nation-wide programme for implementing formative assessment. Achievement is reported to students and parents about twice a year in terms of levels, but there is no central collection of these results and they are not used for national monitoring or for creating league tables of schools (although school results are published for the external testing at age 15–16). Individual student assessment is based on teachers' judgments, moderated by the use of external tests administered by teachers when they judge students as able to pass a test at a certain level. National monitoring is conducted through a separate programme of testing which involves samples of pupils at certain ages.

By contrast, in England a combination of national testing and teachers' judgments is used for internal summative assessment and national test results are used for monitoring performance of students at age 7, 11 and 14 year on year. National test results are also to evaluate the performance of schools, local authorities and the country as a whole. Test and examinations results are also used to create targets for schools and give rise to league tables. (There is more detail on these and other systems in Chapters 6 and 7.)

Differences between the ways in which certain components of a system are carried out matter. Changes made in one part of a system have implications for how other parts can function. It is not difficult to point to examples of this interaction. For example, there is research evidence that

- when school accountability is based on external summative assessment data, this impacts on the way that teachers conduct their own internal summative assessment and on how they use assessment formatively (see for example Pollard et al., 2000);
- how teachers carry out formative assessment can influence their own internal summative assessment practice (Black et al., 2003);
- how moderation of teachers' judgments is carried out can affect the evidence they gather and report about students' achievements (Donnelly et al., 1993).

Often unwanted effects in one part of the system on another arise from piecemeal changes in policy that ignore the relationship of parts within the whole. These interactions, and those between assessment and the curriculum and teaching methods (pedagogy), are indeed at the heart of

arguments made here for change in systems where student assessment is largely based on tests.

Before embarking on making these arguments, however, we take a closer look at the components and clarify some of the terms used in describing the various ways in which these can be carried out.

Components of an Assessment System

The components of a system can be described in terms of combinations of the variables set out in Figure 1.1 under seven headings: purpose, use, type of task that provides evidence, who makes the judgment, the basis for judging the evidence, the way in which the results are reported and any moderation procedures that are needed. Some of the main variables in relation to these seven aspects are also indicated in Figure 1.1. (A far more detailed taxonomy of just one type of assessment, summative assessment by teachers, was developed by Wilmut, 2004.)

The variables in Figure 1.1 can be used to describe whole systems, as well as particular types of assessment within them. For example, the profile of these aspects for formative assessment (assessment for learning) would be:

Purpose formative
Use helping learning
Type of task regular work and some tests/tasks created by the teacher
Agent of judgment teacher and students combined
Basis of judgment criterion-referenced (detailed and task specific) and student-referenced
Form of report or feedback comment or oral feedback.

Moderation does not apply to formative assessment, but in the case of the focus of this book – assessment of learning – all seven headings do apply and Figure 1.1 makes clear the very large number of ways in which assessment for this purpose can be carried out using different combinations of the variables. Not surprisingly, how this is done matters and has a considerable impact on those involved, quite apart from any reaction to the result of the assessment. The choice of variables and how they are combined has implications for what evidence is included (validity), the accuracy of the result (reliability), the impact on those involved, and the cost of the operation in time and other resources. These four properties will be briefly discussed later in this chapter, but also revisited throughout the book as they underpin the main arguments for reform in assessment policy.

First we look briefly at each of the seven aspects in Figure 1.1 and in particular at the different ways in which they can be put into practice in summative assessment.

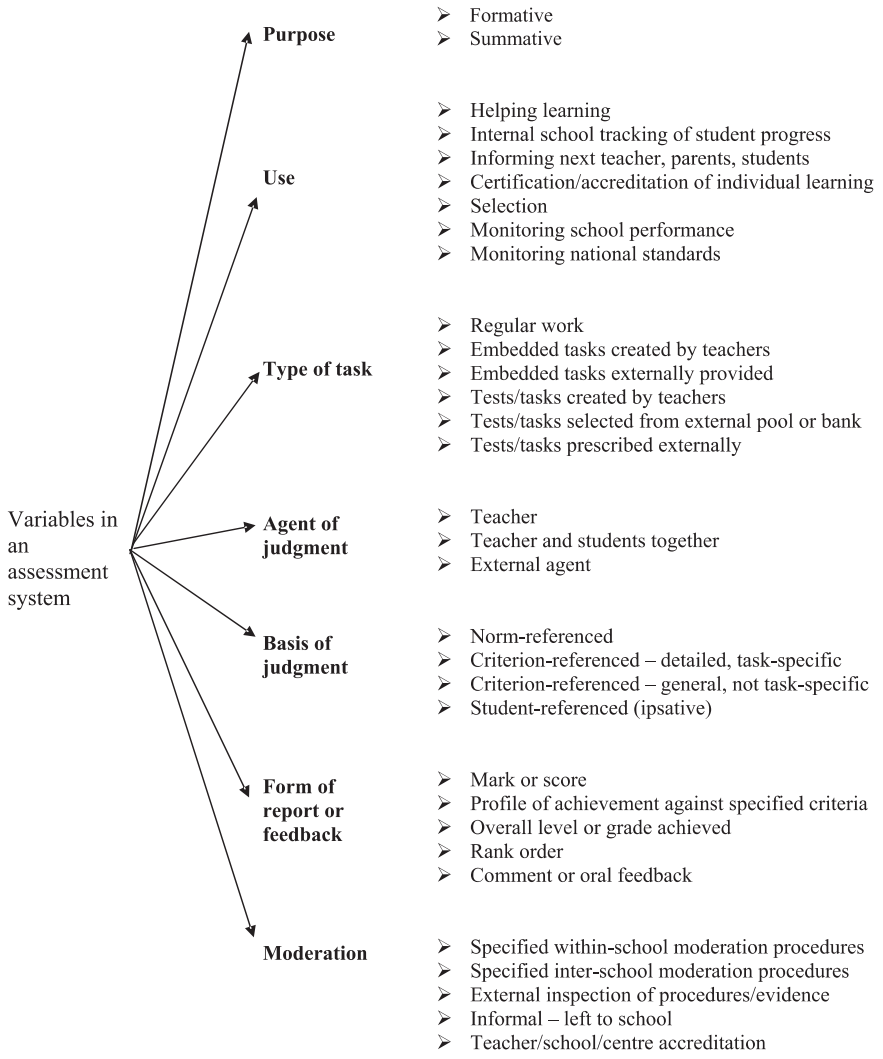

Figure 1.1 Components and variables of an assessment system

Purposes of Assessment

There are two main purposes for assessing students: to inform decisions about learning experiences and to report on what has been achieved. 'Formative' means that the assessment is carried out in order to help learning. It is detailed and relates to specific learning goals. It is essentially part of an approach to teaching and learning in which information about what students have achieved is used to inform decisions as to how to make progress. For this reason it is also called 'assessment *for* learning' and although some-

times a distinction between 'formative' and 'assessment *for* learning' is forged (Black et al., 2002), both terms are widely used as meaning:

> the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there. (ARG, 2002b)

Summative assessment is carried out for the purpose of reporting the achievement of individual students at a particular time. It relates to broader learning goals that can be achieved over a period of time. It can be conducted by giving a test or examination at that time, or summarizing achievement across a period of time up to the reporting date. Each of these approaches can take a number of different forms and their relative pros and cons, impacts and costs are explored in some detail later.

Although the main focus here is on summative assessment (assessment *of* learning), we shall also look at the relationship between formative and summative assessment. This includes examining the extent to which there is evidence to support the different claims that 'Any attempt to use formative assessment for summative purposes will impair its formative role' (Gipps, 1994: 14) or the conclusion that there are ways of using summative tests formatively (Black et al., 2003). The reason for concern about how these two purposes relate to each other is the evidence that summative assessment, particularly when conducted through testing and when there are high stakes attached to the results, can inhibit the use of assessment formatively. Indeed, this concern was a major reason for bringing together evidence about the properties and impacts of summative assessment conducted in different ways. Since both formative and summative purposes are important in education – it is not a matter of 'formative, good', 'summative, bad' – it is essential to discuss how summative assessment can be conducted most effectively and without negative consequences for formative assessment. (A more detailed discussion of these matters is in Chapter 8.)

Use of Assessment Results

For formative assessment there is one main use for the data – to help learning. Indeed, this use defines formative assessment and if the information about students' learning is not used to help that learning, then the process cannot be described as formative assessment. By contrast, the data from summative assessment are used in several ways, some relating to individual students and some to the aggregated results of groups of students.

For individual students, the uses include reporting to parents, other teachers, tracking progress, and selection, certification or accreditation by an external body. These can be grouped under two main headings of 'Internal' and 'External' to the school community:

- 'Internal' uses include using regular grading, record keeping, informing decisions about what courses to follow where there are options within the school, reporting to parents and to the students themselves.
- 'External' uses include certification by examination bodies or for vocational qualifications, and selection for further or higher education.

In addition to these uses, which relate to making judgments about individual students, results for groups of students are used in evaluating the effectiveness of the educational provision for students. The main uses of aggregated results are

- Accountability – for the evaluation of teachers, schools, and local authorities, although the extent to which this should depend on measures of students' achievement is problematic.
- Monitoring – within and across schools in particular areas and across a whole system for year-on-year comparison of students' average achievements, but whether aggregated individual data are the most useful is contested.

In both cases, there are issues about using the results of individual students in these ways. Information yielded is restricted and does not meet the needs of users. (We return to these matters in Chapter 9.)

Type of Task

In theory, anything that a student does provides evidence of some ability or attribute that is required in doing it. So the regular work that students do in school is a rich source of evidence about the abilities and attributes that the school aims to help students develop. This evidence is, however, unstructured and varies to some degree from class to class, or even student to student. These differences can lead to unfairness in judgments unless assessment procedures ensure that the judgments made are comparable and equivalent work is assessed in the same way. One way to avoid this problem entirely is to create the same conditions and tasks for all students, that is, to use tests.

Testing is a method of assessment in which procedures, such as the task to be undertaken and often the conditions and timing, are specified. Usually tests are marked (scored) using a prescribed scheme (protocol), either by the students' teacher or external markers (often teachers from other schools). The reason for uniform procedures is to allow comparability between the results of all students, who may take the tests in different places. Just as assessment can be qualified according to the type of task, so tests are described as 'performance', 'practical', 'paper-and-pencil', 'multiple choice', 'open book', and so on. More formal tests, leading to a certificate or qualification, are often described as examinations.

Teachers regularly create their own tests for internal school use; in other cases they are created by an agency external to the school. Tests are criticised on a number of points, considered in more detail in Chapter 4, but it is the emotional reaction of many students to them that is a considerable cause of concern. The test items are unknown and students have to work under the pressure of time allowed. This increases the fear that they will 'forget everything' when faced with the test; the anticipation is often as unpleasant as the test itself. To counter this, and also to assess domains that are not adequately assessed in written, timed tests or examinations, assessment tasks may be embedded in normal work. The intention is that these tasks are treated as normal work. It may work well where the use is internal to the school, but the expectation of 'normality' is defeated when the results are used for making important decisions and the tasks become the focus of special attention by teacher and students. One example here is the coursework element of GCSE examinations, which has been the subject of critical review (see the example at the end of this chapter and Chapter 6).

Since tests and examinations are forms of assessment, the broader term, 'assessment', is used here when referring to all forms. There is a problem, however, in distinguishing testing and examinations from other methods of assessment, such as on-going assessment by teachers. Some authors attempt to identify assessment as different from testing but then fall into the trap of referring to assessment as a general category, as in 'assessment in schools', thus meaning all kinds including tests. To avoid this, the term 'assessment' will be qualified as 'assessment by testing' or 'teachers' on-going assessment' where there is any risk of ambiguity.

Agent of Judgment

The role that teachers can take in summative assessment varies from collecting evidence required and marked by external agencies (as in administering external tests and examinations) to themselves collecting evidence of, and making judgments about, the achievements of their own students. In between there are various ways in which teachers can be involved, for example in writing and marking examinations, moderating others' judgments and collecting evidence as prescribed by external agencies that is then assessed by others. Consequently the term 'teachers' assessment' can be interpreted in different ways – and the term 'teacher assessment' is even more ambiguous.

The definition of teachers' assessment used here is

> The process by which teachers gather evidence in a planned and systematic way in order to draw inferences about their students' learning, based on their professional judgement, and to report at a particular time on their students' achievements. (ARG, 2006: 4)

This draws attention to the process being no less 'planned and systematic' than alternatives using tests or tasks. Judgments have to be supported by evidence which can be gathered over time from regular work, accumulated in the form of a portfolio of particular kinds of work, or intermittently from project reports, or from specified assignments, or from occasional tasks such as fieldwork or presentations. The reference to 'their own students' emphasises that the meaning is restricted to those situations where teachers assess their own students and exclude the marking of other students' work (apart, that is, from what is involved in moderating other teachers' judgments).

Basis of Judgment

Making a judgment in assessment is a process in which evidence is compared with some standard. The standard might be what others (of the same age or experience) can do. This is norm-referencing and the judgment will depend on what others do as well as what the individual being assessed does. In criterion-referencing the standard is a description of certain kinds of performance and the judgment does not depend on what others do, only on how an individual's performance matches up to the criteria specified. In student-referenced or ipsative assessment a student's previous performance is taken into account and the judgment reflects progress as well as the point reached. The judgments made of different students' achievements are then based on different standards, which is appropriate when the purpose is to help learning but not appropriate for summative purposes.

Summative assessment is either criterion-referenced or norm-referenced. Criterion-referencing is intended to indicate what students, described as having reached a certain level, can do. But using criteria is not a straightforward matter of relating evidence to description (Wilmut, 2004). For example, the level descriptions of the National Curriculum assessment comprise a series of general statements that can be applied to a range of content and context in a subject area. Not all criteria will apply to work conducted over a particular period and there will be inconsistencies in students' performance – meeting some criteria at one level but not those at a lower level, for instance. Typically the process of using criteria involves going to and fro between the statements and the evidence and some trade-off between criteria at different levels, all of which involve some value judgments. Agreement among practitioners will then depend on the extent of shared experience, understanding and values. There may also be some comparison of the work of different students, thus introducing an element of norm-referencing in the interpretation of criteria.

Whilst 'best fit' is used in the National Curriculum assessment and in the teacher-assessed elements of the GCSE, in vocational education the student must meet every criterion in order to obtain an award at a particular level.

In the interests of clear-cut judgments the criteria are detailed, which tends to lead to a somewhat mechanistic approach to assessment, and to learning that is shallowly directed at 'ticking off' particular criteria, rather than adopting a more holistic approach (see Chapter 6).

Form of Report

The form of report depends to a large extent on the nature of the task, the basis for judgment and the audience for the report. Numerical scores from tests are a summation over a diverse set of questions and so have little meaning for what students actually know or can do for the same total can be made up in many ways. Scores also give a spurious impression of precision, which is very far from being the case (see Chapter 4). Converting scores to levels or grades avoids this to a certain extent, and also serves to equalize the meaning of a level from year to year in examinations such as the GCSE and national tests, where new items are created each year. When scores change from year to year, this could be because of differences in the difficulty of items or changes in the students' achievements. The boundary scores for grades or levels can be adjusted to account for differences in item difficulty, but this introduces a severe problem of how to decide the 'correct' cut-off score between one grade and the next. Ultimately this depends upon judgment and there is a variety of procedures (often hotly contested) for making these judgments. Black (1998) gives a useful account and critique of some of these methods and their consequences in the national debates about standards.

Scores can be used directly to rank order students, but this is really only useful in the context of selection. Again, a position in a rank order gives no indication of meaning in terms of learning. In theory, reporting against criteria which describe levels or grades can do this, but since a single overall grade or level would have to combine so many different domains as to make it almost meaningless a profile is needed. The shorthand of levels, as used in reporting National Curriculum assessment can be useful for in-school reporting, but for reporting to parents and students the levels need to be explained or replaced by accounts of what students can do.

Moderation

Moderation is generally associated with assessment where teachers make judgments of students' work, although in its wider meaning it has a role in those parts of all types of assessment where decisions are made and have to be checked. Its purpose can be quality control, or quality assurance, or both. There are several different approaches, each with advantages and disadvantages. To anticipate the discussion in Chapter 5, the main methods

for quality control are statistical moderation, inspection of samples, external examinations appeals and group moderation. For quality assurance of the process rather than adjustment of the outcome, methods include defining criteria, providing exemplification, accrediting schools, centres or individuals, visits by verifiers and group moderation. In the case of teachers' assessment, as defined above, the purpose is to align the judgments of different teachers and the most appropriate methods are those that address both quality control and assurance.

The rigour of the moderation process that is necessary in a particular case depends on the 'stakes' attached to the results. Where the stakes are relatively low, as in internal uses of summative assessment, within-school moderation meetings are adequate, whilst inter-school meetings are needed when the results are used for external purposes. However, the use of exemplification can be seen as a substitute for moderation meetings, thus reducing opportunities for inter-school discussions and for the professional development that these meetings can provide.

Properties of Assessment

Whilst the components and variables set out in the last section offer a means of analytic description of assessment procedures and systems, in order to judge how effective they are for their purpose they need to be evaluated in terms of required properties.

One obvious property is that any assessment should be valid for its purpose, that it assesses what it is intended to assess. Another is that it should provide reliable or trustworthy data. But there are also other matters to be taken into account; in particular, and in view of the interdependence of the various system components, the impact on other assessment practices and on the curriculum and pedagogy. Further, there is the use of resources; assessment can be costly, both in terms of monetary resources and students' and teachers' time.

Thus assessment for any purpose can be evaluated in terms of the extent to which it meets the requirements of its uses for validity and reliability, positive impact, and good use of resources. As we will be referring to these properties frequently throughout the book we consider them briefly here, first individually and then looking at the interactions among them.

Validity

In this context validity means how well what is being assessed corresponds with the behaviour or learning outcomes that it is intended should be assessed. This is often referred to as 'construct validity'. Various types of validity have been identified, most relating to the type of evidence used in

judging it (for example, face, concurrent, content validity), but there is general agreement that these are contained within the overarching concept of construct validity (Messick, 1989; Gipps, 1994). The important requirement is that the assessment concerns all aspects – but only those aspects – of students' achievement relevant to the particular purpose of that assessment. Including irrelevant aspects is as much a threat to validity as omitting relevant aspects. Thus a clear definition of the domain being assessed is required, as is adherence to it.

Reliability

The reliability of an assessment refers to the extent to which the results can be said to be of acceptable consistency or accuracy for a particular use. This may not be the case if, for instance, the results are influenced by whoever conducts the assessment or they depend on the particular occasion or circumstances at a certain time. Thus reliability is often defined as and measured by the extent to which an assessment, if repeated, would give the same result.

The degree of reliability necessary depends on the purpose and use of an assessment. When assessment is used formatively, it involves only students and teachers. No judgment of grade or level is involved, only the judgment of how to help a student take the next steps in learning, so reliability is not an issue (see Chapter 8). Information is gathered frequently by teachers who will be able to use feedback to the student to correct any mistaken judgment. However, high reliability *is* necessary when the results are used by others and when students are being compared or selected.

Impact

Here impact means the consequences of the assessment, often referred to as 'consequential validity' (Messick, 1989). It concerns the inferences drawn from the assessment information in relation to the uses that are made of it. As noted earlier, assessment generally has an impact on the curriculum and on pedagogy, which is greater the higher the stakes attached to the outcomes of assessment, so it is important that any potential adverse effects are minimized. Assessment can only serve its intended purpose effectively if this is the case. The impact is likely to be greater the more frequently summative assessment is carried out. Often teachers mistakenly assume that more summative assessment is needed than is actually the case. In particular, when there are external tests many teachers will feel under pressure to spend time preparing and practising for them, thus making what ought to be an infrequent occurrence into a frequent one.

A key factor in determining the degree and nature of the impact of student assessment is the use of results for the evaluation of teachers,

schools and local authorities. The evidence for this is considered in Chapter 9 which also suggests how the most serious impacts can be avoided.

Resources

The resources required to provide an assessment ought to be commensurate with the value of the information to users of the data. The resources may be teachers' time, expertise and the cost both to the school and to external bodies involved in the assessment. In general there has to be a compromise, particularly where a high degree of reliability is required. There is a limit to the time and expertise that can be used in developing and operating, for example, a highly reliable external test or examination. Triple marking of all test papers would clearly bring greater confidence in the results; observers visiting all candidates would increase the range of outcomes that can be assessed externally; training all teachers to be expert assessors would have great advantages – but all of these are unrealistic in practice. Balancing costs and benefits raises issues of values as well as of technical possibilities.

Interaction Among the Properties

As just noted, the extent to which the property of reliability can be optimized is limited in practice by resources. Similarly, changes in procedures aimed at increasing validity, say by doubling the time used in testing, would increase the impact on the resource of teaching and learning time. A less obvious but key interaction is between reliability and validity. In essence it means that, in practice, an assessment cannot have both high validity and high reliability. This applies to whatever way an assessment is carried out.

Take tests, for example. No test can cover all the learning that is set out in the curriculum. What is tested can only be a sample of the curriculum goals and in order to make the test as reliable as possible, the sample will inevitably be biased towards those aspects that can be consistently marked or marked by machine. This favours items assessing factual knowledge and the use of a closed item format, as opposed to items requiring application of knowledge and the use of more open-ended tasks. The consequent limitation on what is covered in a test affects its validity; increasing reliability decreases validity. Attempts to increase validity by widening the range of items, say by including more open-response items where more judgment is needed in marking, will mean that the reliability is reduced.

The same arguments apply to the use of teachers' judgments instead of tests for summative assessment. Whilst validity can be high, since the data can include all outcomes, reliability will be low unless effective moderation procedures are applied (see Chapter 5). Attempts to increase reliability by

standardizing the tasks that are assessed by teachers lead to narrow, artificial tasks of low validity. Black et al. (2004) describe a stark example of this in science:

> It is ironic that the only aspect of science that is entrusted, at GCSE level, to teachers' assessment has led to 'investigations' which the various external pressures have reduced to stereotyped exercises that are widely recognised to be of no interest to students and to present them with a mockery of scientific enquiry. Similar damaging effects of moderation that lead to 'rubric-driven instruction' have been reported in other subjects and in other countries. (Paechter, 1995; Baker and O'Neil, 1994) (Black et al., 2004: 5).

In recognition of the interaction of validity and reliability it is sometimes useful to refer to the combination of the two as *dependability*. The definition of this term used here gives priority to validity, so that it is taken to mean 'the extent to which reliability is optimised while ensuring validity'.

2 | Assessment and the curriculum

The chapter begins by considering what information we want assessment to provide about students' learning. We note a considerable amount of official support for changes in the curriculum that would better provide for the needs of students in a world that is rapidly changing. The arguments focus on the importance of helping students to develop various 'literacies' – meaning a broad understanding of concepts in each area that enables effective engagement in modern life – such as creativity and economic productivity, citizenship, learning with understanding and learning how to learn.

We argue that the absence of representation of these goals in the information provided by many current assessment systems is partly to blame for inhibiting the real changes in educational practice that are needed. Yet the assessment of these important goals is possible and some brief examples have been indicated. The major change, however, is to move away from traditional assessment methods based on tests and to make more use of teachers' judgments. Evidence and arguments for this course of action are discussed in Chapter 4.

Introduction

What is assessed influences what is taught and how it is taught, and hence the opportunities for learning. The relationship between assessment, the curriculum and teaching methods (pedagogy) is often represented as a triangle, as in Figure 2.1, to show that each one of these has some relationship with the other two.

Of course this simple representation does not indicate the direction of the effect of one feature on another. Does the assessment influence the curriculum or the curriculum influence the assessment? Is the curriculum that students experience the same as the intended curriculum? Similar questions apply to assessment and pedagogy, for what teachers do in the classroom is

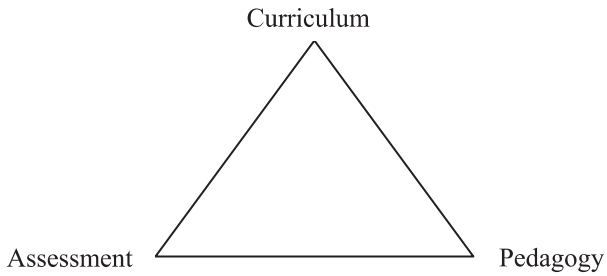

Figure 2.1 The interactions between assessment, curriculum and pedagogy

not always what they would like to be able to do (James and Pedder, 2006). Ideally we would like to think of assessment and pedagogy following the curriculum, so that methods of teaching and the assessment of outcomes are chosen to be appropriate to what we want our students to learn. Unfortunately, there are several reasons why we do not find this in reality and this can have serious implications for students' learning.

Some of these reasons follow from the uses made of the results, especially when these include judgments that affect the future of students or the status of teachers, schools and even countries, giving the assessment results what has been described as 'high stakes'. Although high stakes can be attached to the results of assessment conducted by teachers, when results are used for evaluation they invariably are derived from tests because of their supposed precision. Thus the testing itself as well as the use of the results increases the impact on the curriculum, on learning and on teaching. We return to these matters in Chapter 9 but in this chapter the focus is on the relationship between assessment and the curriculum; what we ought to assess and whether current arrangements enable us to do it.

The concern is with the substance, not the form of the assessment, yet there is evidence that the form of traditional methods continues to restrict what is assessed. The first consideration ought to be about what we want to assess and this will be the starting point here. There is ample support for new goals in education related to preparing young people for life in modern society. Yet when we come to ask whether these goals are reflected in students' school experience, the answer is not encouraging. One important reason for this is that they are not being assessed. The second part of the chapter looks at some of the difficulties of assessing these goals and what is needed to overcome these problems. This is the start of making a case for greater use of teachers' assessment.

What We Ought to Assess

A key question is, how consistent are current assessment procedures and content with a curriculum that is needed to prepare students for modern life? Current thinking, world-wide, emphasises the importance of developing in students types of skill, attitudes, knowledge and understanding that are regarded as more important than accumulating large amounts of factual knowledge. Content knowledge can be found readily from the information sources widely available through the use of computers and especially the internet. What are needed are the skills to access these sources and the understanding to select what is relevant and to make sense of it. So we need to provide students with an understanding of broad, widely applicable concepts and the ability to use them to solve problems and make decisions in new situations. This is often expressed in terms of becoming 'literate' in relation to different subject areas.

Facing new problems will be the normal run of things rather than the exception in the future lives of young people, as changes in technology affecting everyday life occur at an increasing rate. Dealing with problems responsibly involves a broad understanding of society and democratic processes and some experience in participating in them. Being able to manage change, and do more than respond to problems, requires creativity and enterprise. Further, underlying all attempts to prepare students for meeting change confidently is the need for them to learn with understanding and to learn how to learn.

It is not new to suggest that these are important outcomes of a modern education. Reference to them abounds in official documents and in reports and recommendations from a number of influential groups and councils. The questions we now need to raise are: how well are these aspects reflected in the way that students' achievements are assessed? What changes would be needed to make sure that they are included and dependably assessed? One thing is sure – if they are not included in the assessment, then the attention they receive will not match the rhetoric of their importance. Before considering how they might be assessed, however, we need to take a closer look at what these aspects are, why they are important goals of learning and what kinds of experiences promote their development.

The Development of 'Literacies'

The notion of being 'literate' has for some time been extended beyond the ability to read and write. The term now generally carries the connotation of being able to engage effectively with different aspects of modern life. So it is common to refer to 'technological literacy', 'mathematical literacy', 'scientific literacy', even 'political and social literacy'. Being literate in these

various respects indicates having the knowledge and skills that are needed by everyone, not just those who will be specialists in or will make a career using knowledge of one of these areas.

The emphasis is not on mastering a body of knowledge but on having, and being able to use, a general understanding of the main or key ideas in making informed decisions and participating in society. Literacy, used in these ways, does not mean reading and writing *about* technology, mathematics, science, and the like. Rather, it means having the understanding and skills to participate in the discussion of issues and to make decisions in the many matters of everyday life that involve technology, science, politics and so on.

What this means in practice is spelled out in relation to scientific literacy by Millar and Osborne (1998), who recommend that 'the science curriculum from 5 to 16 should be seen primarily as a course to enhance general "scientific literacy"' (p. 4). They explain that this means

> that school science education should aim to produce a populace who are comfortable, competent and confident with scientific and technical matters and artefacts. The science curriculum should provide sufficient scientific knowledge and understanding to enable students to read simple newspaper articles about science, and to follow TV programmes on new advances in science with interest. Such an education should enable them to express an opinion on important social and ethical issues with which they will increasingly be confronted. It will also form a viable basis, should the need arise, for retraining in work related to science or technology in their later careers. (Millar and Osborne, 1998: 4)

Some empirical support for this view of what should be taught in science was provided by one of four research projects undertaken by the Evidence-based Practice in Science Education (EPSE) Research Network, part of the ESRC's (Economic and Social Research Council) Teaching and Learning Research Programme. Twenty-five 'experts', representing the main 'stakeholder' groups in science education, took part in a three-stage Delphi study. The outcome showed considerable agreement among all groups on the most highly rated themes (Millar et al., 2006). These were grouped under three headings:

- Nature of scientific knowledge (tentative nature of science; historical development);
- Methods of science (hypothesis and prediction; diversity of scientific thinking; creativity; science and questioning);
- Institutions and social practices in science (co-operation and collaboration in development of scientific knowledge).

A similarly broad and applied view underpins the OECD's PISA (Programme of International Student Assessment) definition of mathematical literacy as

> An individual's capacity to understand the role that mathematics plays in the world, to make well-founded mathematical judgments and to engage in mathematics, in ways that meet the needs of that individual's current and future life as a constructive, concerned and reflective citizen. (OECD, 1999: 41).

It is not difficult to see that working with such goals in mind is different from seeing science and mathematics (and history, politics and economics) as a succession of facts or algorithms to be learnt, with no attention to the overarching coherence between concepts and their relevance in students' current and future lives.

Creativity and Economic Productivity

In 2001 the Department of Trade and Industry (DTI) joined with the Department for Education and Employment (DfEE) in England in producing a paper entitled 'Opportunity for All in a World of Change'. It made a case for school education giving more attention to development of the skills and attitude needed in business and industry.

> People who generate bright ideas and have the practical abilities to turn them into successful products and services are vital not just to the creative industries but to every sector of business. Our whole approach to what and how we learn, from the earliest stages of learning, needs to adapt and change in response to this need. Academic achievement remains essential but it must be increasingly delivered through a rounded education which fosters creativity, enterprise and innovation … (DTI and DfEE, 2001: para 2.11)

The emphasis on the need to change 'what and how we learn' underlines the links between curriculum and pedagogy, for qualities such as 'creativity' cannot be 'taught' in the same way as, say, the 'basics' of reading and arithmetic. Rather, creativity, innovation and enterprise are best engendered in schools when teachers and management practise them in their planning and provision (Hargreaves, 2003). Instead of following 'top-down' approaches, such schools develop and test out methods of providing opportunities for students and teachers to discuss, collaborate, link with other schools and local industry and use technology in teaching and learning.

Citizenship

Some of the aims, and ways of working towards them, to develop citizenship overlap with those just discussed. Citizenship includes social and moral responsibility, community involvement and political literacy, according to the DfES's Citizenship website. Again, the ways of working needed to help students achieve these aims mean enabling them to participate in, and

not just learn about, links with the local community and the wider world outside school. These require students not only to learn about their neighbourhood and social and political structures, but to become involved in service to the community and to take responsible care of the environment. Through this they develop 'self-confidence and socially and morally responsible behaviour both in and beyond the classroom, towards those in authority and towards each other' (DfES, Citizenship website).

Learning with Understanding

Understanding occurs frequently in the expression of learning goals, but rarely is its meaning clear and the complexity of the concept acknowledged. Taken seriously, the development of understanding has strong implications for curriculum content and pedagogy, as do the aims of developing enterprise and citizenship. Understanding is quite different from knowing facts, although it requires factual knowledge, as White points out:

> [understanding] is a continuous function of a person's knowledge, is not a dichotomy and is not linear in extent. To say whether someone understands is a subjective judgement which varies with the judge and with the status of the person who is being judged. Knowledge varies in its relevance to understanding, but this relevance is also a subjective judgment. (White, 1988: 52)

Different dimensions and levels of understanding also have been identified, for example by Wiske who considers three dimensions related to the form of communication and four levels of depth of understanding: naïve, novice, apprentice and master (Wiske, 1998: 180). Understanding shows in the ability to organize knowledge, to relate it actively to new and to past experience, forming 'big' ideas, much in the way that distinguishes 'experts' from 'novices' (Bransford et al., 1999). Big ideas are ones that can be applied in different contexts; they enable learners to understand a wide range of phenomena by identifying the essential links ('meaningful patterns' as Bransford et al. put it) between different situations without being diverted by superficial features. Merely memorizing facts or a fixed set of procedures does not support this ability to apply learning in contexts beyond those in which it was learned. Knowledge that is understood is thus useful knowledge that can be used in problem solving and decision making.

Learning How to Learn

An important part of preparing young people for life and work in the rapidly changing society of today and tomorrow is to help them develop awareness and understanding of the process of learning – a key aspect of meta-cognition. Throughout their lives they will have to make more

choices than those who lived in past decades and both work and leisure will involve activities that we can as yet only guess at. This is underlined by the OECD, who point out that:

> Students cannot learn in school everything they will need to know in adult life. What they must acquire is the prerequisites for successful learning in future life. These prerequisites are of both a cognitive and a motivational nature. Students must become able to organise and regulate their own learning, to learn independently and in groups, and to overcome difficulties in the learning process. This requires them to be aware of their own thinking processes and learning strategies and methods. (OECD, 1999: 9).

The ability to continue learning throughout life is acknowledged as essential for future generations and thus it has to be a feature in the education of every student. Since learners have to be prepared for meeting the challenge of learning anew throughout their lives, they need to learn how to learn. *Learning how to learn* is not the result of being taught to use a set of higher-order skills, but rather of having used a set of effective learning practices and applied them in various contexts (Black et al., 2006a). It is important that learning how to learn is seen as integral to, and a consequence of, effective learning. What is required for understanding learning is, therefore, no more than helping students to think about and reflect on their learning as part of the learning process. Thus meta-cognition is seen as being embedded in learning processes and is developed, as with other learning, through interaction and discussion with other students and the teacher. Learning collaboratively provides students with feedback and scaffolding that supports their understanding of learning as well as requiring and developing other skills related to problem solving and communication. Hargreaves (2007) suggests that collaborative learning and its assessment improve performance in traditional examinations.

Learning about how to learn and the ability to reflect on the adequacy of what one knows is the key to taking steps towards further learning. Research, such as that reviewed by Black and Wiliam (1998a), shows that the ability to take effective action results from students being helped to:

- see how to improve their work, by feedback that is non-judgmental;
- try to explain things rather than just describe them;
- take some responsibility for assessing their own work, finding the errors in their own or a partner's work;
- talk about and justify their reasoning;
- understand the goals and the quality of work they should be aiming for.

These are key features of using assessment to help learning (formative assessment or assessment for learning). It follows that an important requirement for summative assessment is that it supports and does not inhibit the practice of formative assessment (see Chapter 8).

Can These Goals Be Assessed?

Returning to the initial statement at the start of this chapter – that what is taught is influenced by what is assessed – raises some unavoidable questions. Are these widely advocated components of a modern education reflected in what and how we assess? If not, why not? There are three points to make in relation to these questions before tackling a further one. Can anything be done about it?

First, we can dismiss any doubt that assessment does influence teaching. More and more research studies are confirming this, the latest being a series carried out by the Learning How to Learn project (James et al., 2006). Evidence from 1,500 staff in 40 primary and secondary schools in England led to the conclusion that there is no doubt that teachers are teaching to the tests their pupils have to take; they feel they cannot do anything else. Case studies revealed that teachers believed that 'there are circumstances beyond their control which inhibit their ability to teach in a way they understand to be good practice' (Marshall and Drummond, 2006: 147).

Second, it is not difficult to see from what is assessed in external tests and examinations the extent to which the skills, understanding and attitudes just discussed are included. The result is 'hardly at all'. Since teachers' internal assessment tends to emulate the external assessment, this also fails to reflect these important goals. There is no suggestion here that only these goals should be assessed, for it is important to know whether students have the knowledge and basic skills that underpin the development of broad concepts and the different forms of literacy. But it is essential that all valued goals are included in assessing students' progress. At present this is not the case in systems such as that in England.

Third, in relation to the reasons for this neglect, some points made in Chapter 1 are relevant at this point. For assessment where the results are used beyond the school, tests are preferred to other forms of assessment because they are considered to be reliable and to be 'fair'. In fact the assumption of 'fairness' is not justified (see Chapter 4), but leaving that aside and focusing for the moment on the measured reliability on items, it was noted (p. 26) that tests are preferred because they are viewed as having higher reliability than other forms. However, tests are only as reliable as the scoring or marking and steps taken to make reliability as high as possible favour items that are closed, so that marking depends as little as possible on human judgment. Clearly, items that require students to be creative or to present arguments or show understanding of a complex situation do not fit this description. Consequently they are rarely considered and would be unlikely to survive the pilot trials used in developing and selecting items for external tests and examinations.

What is Possible

It is legitimate to ask at this point whether it is indeed possible to create test items that assess application, problem solving, critical thinking, and so on. Perhaps surprisingly the answer is positive – but with a caveat. Some items of these kinds were included in surveys conducted nationally, by the Assessment of Performance Unit (APU) in England, Wales and Northern Ireland in the 1980s, and such items currently feature in the National Assessment of Educational Progress (NAEP) in the United States. International surveys of the OECD's PISA are wholly concerned with the assessment of scientific, mathematics and reading literacy. For example, the item in Figure 2.2 was created as part of a bank to assess scientific literacy by the PISA.

In the APU, the skills of enquiry were assessed through individual practical investigations as in the example in Figure 2.3.

These examples, dealing with real things or real data, are highly dependent on the choice of content. The surveys in which they are used provide the evidence that students who perform well in one item will not necessarily do so in another item testing the same skills but in a different context. In a recent carefully-designed study in the USA, Pine et al. (2006) assessed fifth grade students using several 'hands-on' performance tasks, including one based on 'Paper Towels' (as in Figure 2.3) and one called 'Spring', about the length of a spring when different weights were hung on it. They found 'essentially no correlation for an individual student's scores. Students with either a 9 or a 1 Spring score had Paper Towels scores ranging from 1 to 9' (Pine et al., 2006: 480). For particular tasks selected from a wide range of possible tasks, the 'task sampling variation', is large and this means that to obtain a reliable score for an individual student would require that individual to tackle a totally unacceptable number of tasks.

For individual students the tasks have high validity but they are low on reliability and fairness, since a student's score is highly dependent on the nature of the tasks chosen. However, in a survey where students are sampled it is indeed possible for a large number of tasks to be given, since any one student takes only a sample of the total items and tasks. The scores of individual students in these surveys are not relevant and become meaningful only when combined with those of other students in the sample.

Although extended performance or practical tasks seem particularly prone to the task sampling variation, there is an element of this problem in every test for individual students since these can only contain a sample of possible items.

Read the following information and answer the questions which follow.

WHAT HUMAN ACTIVITIES CONTRIBUTE TO CLIMATE CHANGE?

The burning of coal, oil and natural gas, as well as deforestation and various agricultural and industrial practices, are altering the composition of the atmosphere and contributing to climate change. These human activities have led to increased concentrations of particles and greenhouse gases in the atmosphere.

The relative importance of the main contributors to temperature change is shown in Figure 1.

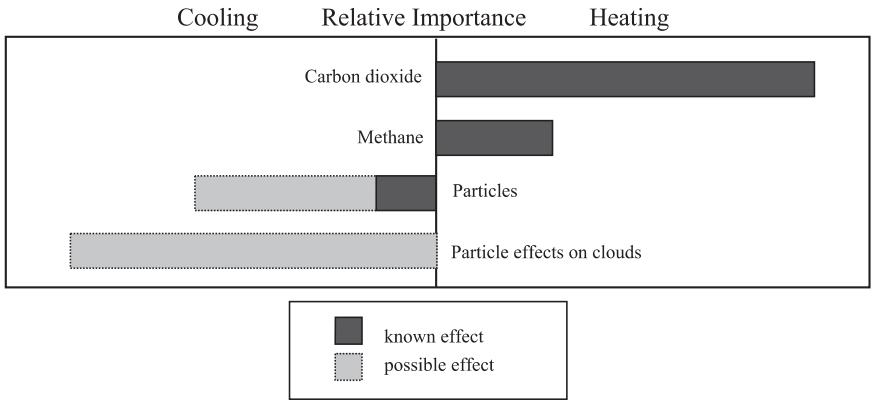

Figure 1: Relative importance of the main contributors to change in temperature of the atmosphere. Source: adapted from http://www.gcrio.org/ipcc/qa/04.html

Bars extending to the right of the centre line indicate a heating effect. Bars extending to the left of the centre line indicate a cooling effect. The relative effect of 'Particles' and 'Particle effects on clouds' are quite uncertain: in each case the possible effect is somewhere in the range shown by the light grey bar.

Figure 1 shows that increased concentrations of carbon dioxide and methane have a heating effect. Increased concentrations of particles have a cooling effect in two ways, labelled 'Particles' and 'Particle effects on clouds'.

Item 1:
Use the information in Figure 1 to support the view that priority should be given to reducing the emission of carbon dioxide from the human activities mentioned.

Item 2:
Use the information in Figure 1 to support the view that the effects of human activity do not constitute a real problem.

Figure 2.2 An item used in the PISA assessment of Scientific Literacy (OECD, 2000)

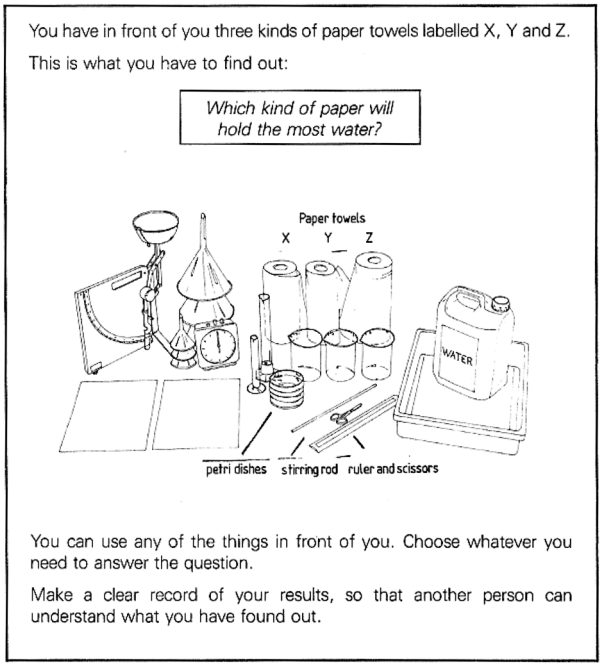

Figure 2.3 The Paper Towels investigation in the APU (Welford et al., 1985)

Alternatives to Tests

Since valid and reliable assessment of some of the achievements that ought to be assessed comes up against some unavoidable obstacles, are there any alternatives for summative assessment? Fortunately there are, all of them depending on the fact that the experience that students need in order to develop the desired skills, understanding and attitudes also provides opportunities for their progress to be assessed. The key factor is the judgment of the teacher. Assessment by teachers can use evidence from regular activities supplemented if necessary by evidence from specially devised tasks introduced to provide opportunities for students to use the skills and understanding to be assessed.

Over the period of time, such as a semester or half year, for which achievement is being reported, students have the opportunity to engage in a number of activities in which a range of attributes can be developed. These same activities provide opportunities for the development to be assessed by the teacher. In other words, the limitation of the restricted time that a test provides does not apply when assessment is teacher-based.

Methods of assessment based on observation during regular work also enable information to be gathered about processes of learning rather than

only about products. Even the kinds of items shown in Figures 2.2 and 2.3 are not capable of assessing qualities such as reflection on the process of learning. This kind of information is useful to those selecting students for advance vocational or academic courses of study, where whether candidates have learned how to learn and are likely to benefit from further study is as important as what they have already learned. Nor is assessing understanding a simple matter, it is likely to require different kinds of evidence in different curriculum areas and at different stages of learning. When students are learning we want assessment to indicate how far they are progressing along the various dimensions of understanding. The notion of 'big' ideas has to be considered in relation to the experience of learners; for younger learners they will not be as 'big' as for older learners. These larger concepts cannot be 'taught' as such, rather they are created by the active participation of the learner. At all stages, therefore, the assessment also needs to encompass evidence of the ability of learners to engage in using and developing the processes of learning.

Using teachers' judgments for summative assessment is not new; it is built into existing practice in the UK for assessing young children and for older students when assessing vocational skills. The Pre-school or Foundation Stage assessment includes personal, social and emotional development, physical development and creative development, as well as development in basic literacy, numeracy and knowledge of the world (see Chapter 6). These assessments serve a formative as well as a summative purpose and are not for external use, but in other countries teachers' assessment is used for important external assessment (see Chapter 7).

Ways of assessing outcomes such as orientation to lifelong learning and understanding are also available. They depend upon self-reporting by students, but since there are no 'correct' and 'incorrect' responses and no scores, there is no sense of being tested. For example, the Effective Lifelong Learning Inventory (ELLI) comprises a number of statements to which learners respond by indicating their agreement or disagreement on a five-point scale. It provides a profile of a student's characteristics relating to willingness and enjoyment of learning. It can be used with students from the age of about eight onwards. The statements have been found to relate to seven dimensions, each of which is an aspect of effective learning: changing and learning, meaning making, critical curiosity, creativity, learning relationships, strategic awareness and resilience. Examples of statements in the inventory are:

'I can feel myself improving as a learner.'

'When I have trouble learning something, I tend to get upset.'

'Talking things through with my friend helps me to learn.'

> 'I like it when I can make connections between new things I am learning and things I already know.' (Adapted from Deakin Crick et al., 2002: 51–2)

The results can be used by teachers both to identify the help that individual students may need to develop their 'learning power' and also to devise learning strategies and create a classroom climate that favours learning.

Student questionnaires have also been used by Black et al. (2006b) to explore attitudes to and views about learning. In their instrument, students respond by marking one point on a six-point scale between pairs of statements. For example, in some cases this is pairs of opposites:

I enjoy learning ( ) ( ) ( ) ( ) ( ) ( ) I don't enjoy learning

and in other cases, non-exclusive pairs:

I like to be told exactly what to do	( ) ( ) ( ) ( ) ( ) ( )	I like to do things where I can use my own ideas

Black et al. (2006b) found inconsistencies when the questionnaire was completed for a second time by the same students. These were considerable for the younger students (ten year olds) in their study, which the authors suggest may have been in part 'an inevitable effect of immature response to questions that address issues and generalizations which are too novel for them' (Black et al., 2006b: 167). This indicates that there are limitations in using self-response methods with younger students when the concepts involved are not easily expressed in simple language.

Hautamäki et al. (2005) have created the Finnish Learning-to-Learn Scales, which have been widely used in research in Finland and also translated into Swedish and English. There are scales assessing cognitive skills and competencies, beliefs, motivational orientations, self-regulation, self-concept in academic areas, self-worth and self-esteem. High values of reliability are reported for these scales. Validity, however, will depend upon the extent to which they predict how students behave when, as adults, they are faced with situations where they have to apply their learning of how to learn.

In the next chapter we will continue to focus on what ought to be assessed by considering what those who receive and use assessment say about what they want to know about students.