



*ASF Working Paper 3*

**Purposes, properties and methods of  
summative assessment**

# **ASF Working Paper 3: Purposes, properties and methods of summative assessment**

## ***Contents***

### **Background: Working Papers 1 and 2**

### **Introduction to Working Paper 3**

#### **1. What outcomes ought we to assess?**

- 1.1 Learning with understanding
- 1.2 Understanding learning
- 1.3 Consequences for summative assessment

#### **2. Properties of assessment**

- 2.1 Validity
- 2.2 Impact
- 2.3 Reliability
- 2.4 Practicability
- 2.5 The relationship between reliability and validity and the concept of dependability

#### **3. Properties of summative assessment**

- 3.1 Summative assessment for internal purposes
- 3.2 Summative assessment for external purposes

#### **4. The properties of summative assessment by teachers**

#### **5. The relationship between summative assessment of students and assessment for other purposes**

- 5.1 Assessment for summative and formative purposes
- 5.2 Assessment for summative and accountability purposes
- 5.3 Assessment for summative and system monitoring purposes

#### **6. Implications for the role of assessment by teachers in a summative assessment system**

**Appendix** Estimated time and direct costs of summative assessment, including teacher-made and external tests and examinations in primary and secondary schools in England

### **References**

## **Background**

The Assessment Systems for the Future project exists to collect and use evidence from research and practice throughout the world to propose assessment systems in which the process and products of summative assessment are used appropriately for various purposes with maximum benefit to students' education. Within this broad mission, the particular focus of the project is the role that assessment by teachers can take in summative assessment, defined as

*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.*

In its first year the project held three invitational seminars to gather evidence from research and practice within and beyond the UK relating to summative assessment by teachers. The programmes, participants and outcomes of these seminars were reported in full on the ARG website<sup>1</sup>. The project has developed principles of summative assessment by teachers, a framework for discussing and analysing methods of summative assessment by teachers and proposals relating to its implementation. Two working papers were produced. Working Paper 1 set out the arguments behind the principles and proposals. Working Paper 2 summarised the evidence from two systematic reviews of research on aspects of using teachers' judgements for summative assessment. These were discussed at three consultation conferences for policy-makers and practitioners held in November and December 2004. Revised drafts of the working papers were produced as a result of the conferences and are available on the ARG website.

Working Paper 3 came out of discussions at Seminar 4, held in January 2005, and Seminar 5 in July 2005. Seminar 4 reviewed the implications of the consultation conferences and Seminar 5 considered the perspectives of participants in and users of assessment systems, namely students, teachers in schools and other educational institutions, parents and employers.

A fourth Working Paper illustrates how the framework and arguments in earlier papers can be applied to some examples of current policy and practice in a range of educational contexts within the UK.

### **Introduction to Working Paper 3**

ASF is concerned with summative assessment and particularly with the role of teachers' assessment in it. We recognise, however, that summative assessment is only one part of a complete assessment system, which encompasses assessment for a range of purposes, the main ones being:

---

<sup>1</sup> [www.assessment-reform-group.org](http://www.assessment-reform-group.org)

With reference to individual students:

Formative – to help learning and foster deeper engagement with it: essentially a pedagogical approach rather than a separate activity added to teaching.

Summative for ‘internal’ uses - for keeping records and giving reports on progress to other teachers, parents and students; the details of how this is done are a matter for school policy.

Summative for ‘external’ uses - including certification, selection and meeting statutory requirements; procedures are largely determined outside the school.

With reference to aggregate data on the performance of groups of students:

Accountability – for evaluation of teachers, schools, local authorities: the type of evidence, how gathered and by whom, and the extent to which it should depend on measures of students’ achievement are problematic, but are matters that are generally determined by national and local policy rather than individual schools.

Monitoring – within and across schools in particular areas and across a whole system: for year on year comparison of students’ average achievements; at the system level the procedures are planned and administered outside the school.

Assessment for the last two of these purposes is not used to make decisions that directly affect individual students, as it is in the first three purposes. However, using the product of summative assessment for accountability and monitoring can, and does, affect students through impact on teaching, the curriculum and resource allocation<sup>2</sup>. This underlines the fact that the system components that serve different purposes are not independent of each other, a point to which we return later (p 7).

This paper sets out our further thinking about what is needed for effective and equitable summative assessment. Before considering how this can be best put into practice, a prior question concerns what we want to assess. What are the learning outcomes that are valued and ought to be included in summative assessment? This question becomes all the more important given the interaction between assessment, what is taught and how it is taught. Summative assessment can all too easily turn what can be exciting and empowering learning experiences into dull routine aimed narrowly at meeting assessment criteria. Spelling out the kind of learning experiences and the outcomes of learning that we value will not necessarily avoid this but will ensure that unwanted negative impacts of summative assessment practice no longer remain hidden. Further, it ensures that any differences between what we ought to be assessing and what is actually assessed are made evident. Once clear about the requirements in terms of learning goals that ought to be included, we can progress to considering the characteristics or properties of assessment that can meet these requirements. This leads to implications for the role of teachers’ assessment in a dependable system of summative assessment.

---

<sup>2</sup> Evidence for this is summarised in Working Paper 2

The paper begins, in part 1, with a statement of, and rationale for, the learning outcomes that an effective summative assessment system should include. Part 2 sets out the desirable properties of assessment practices and outcomes, which apply to assessment for all purposes. Part 3 then considers how these properties may be incorporated in various methods of assessment. Part 4 provides a summary of the properties of summative assessment by teachers compared with the use of externally devised tests. In part 5 we look at the interaction of different ways of conducting summative assessment on assessment for other purposes – formative, accountability and system monitoring. Finally we bring together the implications for summative assessment by teachers that are noted throughout the paper.

## **1. What outcomes ought we to assess?**

Assessment, the curriculum and pedagogy are inextricably linked. Learning outcomes depend on all three but the driving force is assessment. That assessment influences what is taught and how it is taught is well established by research (reviewed, for instance, by Crooks (1988), Black and Wiliam (1998) and Harlen and Deakin Crick (2003)). It follows that the degree to which an assessment supports the learning goals that we value, should be foremost in deciding its focus and form.

In the case of summative assessment, the concern of ASF, this means that the assessment should give a valid account of learning related to these goals and that it should have no adverse consequences for achieving the goals. It should not, for instance, have the effect that greater attention is given to some goals rather than others purely because these are assessed; neither should it lead to teaching strategies being chosen because they are a short cut to meeting success requirements. In other words, assessment methods should facilitate the assessment of the full range of goals, so that what is assessed is synonymous with the goals of learning.

It is important, then, to take further a discussion, begun in Working Paper 1, about what we mean by ‘the full range of goals’ and why it is that we value them. In all areas of learning, understanding is a key aim. So also is the development of meta-cognition, the capacity to think about one’s learning, about what learning is and how to do it. We need to define and justify these goals of learning before looking at how they can be most dependably assessed.

### **1.1 Learning with understanding**

Why we value understanding seems obvious at first glance but the reasons are, of course, dependent upon what we mean by it. Understanding frequently occurs in the expression of learning goals, but rarely is its complexity acknowledged. It is quite different from knowing facts, although it requires factual knowledge, as White (1988) 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. (p 52).

Different dimensions and levels of understanding have also been identified, for example by Wiske (1998) who considers three dimensions related to the form of communication and four levels of depth of understanding: naïve, novice, apprentice and master (Wiske, 1998, p180). It is not necessary to go into further details, however, to appreciate that assessing understanding is not a simple matter and is likely to require different kinds of evidence in different curriculum areas and at different stages of learning. Understanding shows in the ability to organise 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 memorising facts or a fixed set of procedures does not support this ability to apply learning to contexts beyond the ones in which it was learned. Knowledge that is understood is thus useful knowledge.

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.

### **1.2 Understanding learning**

We value the development of students’ awareness and understanding of the process of learning – the development of meta-cognition - because we recognise the need to prepare young people for life and work in the rapidly changing society of today and tomorrow. Throughout their lives they will have to make more choices than did those living in past decades and both work and leisure will involve activities that we do not now even know about. This is underlined by the OECD, who point out that students cannot learn in school everything they will need to know in adult life. They must acquire in school the cognitive skills and attitudes for successful learning in future life. ‘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, p9).

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* implies that one has practised and can apply a set of effective learning practices. The alternative phrase *‘learning to learn’* implies that there exists a single unitary skill, a claim which seems impossible to justify. We regard it as important that learning how to learn is seen as integral to, and a consequence of, effective learning and not as a set of higher-order skills that can be taught and assessed separately. 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 embedded in learning processes and is developed, as 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.

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 (1998), 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-judgemental;
- 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.

### **1.3 Consequences for summative assessment**

The arguments above raise some fundamental questions in relation to summative assessment. We have argued that assessment should include the processes of learning, yet summative assessment is essentially concerned with outcomes of learning. So, questions arise such as: whilst the value of students' self- and peer-assessment is recognised, should summative assessment be concerned with finding out the extent to which students are able to do this, or only with the learning that it leads to? Similarly, is it possible, or necessary, to identify understanding that is developed through collaboration, through students taking greater responsibility for their learning, and 'learning how to learn'? Or are we only interested in whether students have developed understanding at appropriate levels, regardless of how they may have arrived at it?

Clearly answers to such questions have important consequences for the selection of methods of assessment that are regarded as adequate and effective. The search for answers to these questions begins with what are regarded as 'outcomes'. If we take seriously the goal of developing skills for lifelong learning, then the development of these skills and related motivation to use them have to be regarded as outcomes that should be assessed. Taking this wider view of outcomes, as encompassing skills and processes of developing understanding, as well as the understanding already achieved, means that summative assessment needs to be concerned with students':

- knowledge of facts and principles necessary for understanding the subject;
- understanding, shown in the ability to apply knowledge and explain in new situations;
- ability to access and assess relevant information from a variety of sources
- interest and motivation to learn more;
- ability to assess and regulate their own learning.

Further implications for assessment follow from recognising that students' engagement in learning (reflected, for example, in the concept of personalised learning) requires that they have some choice in the process of learning and experience a curriculum that can be tailored to their own needs. Ideally this allows teachers and students to find the best way to achieve learning goals rather than only through a uniform programme dictated by tests and examinations. Such flexibility is inseparable from embracing goals such as learning with understanding and understanding learning, as suggested above.

Finding ways to assess these outcomes is crucial since outcomes of learning are widely used as indicators of the quality of educational experience both within and across schools. Current indications that this will continue and be based on more detailed data about outcomes (micro-level performance data) make it essential that the data reflect all the important outcomes such as those above and not the few that are conventionally and easily assessed. We therefore need to look critically at what is assessed. If the required outcomes are not being included, then we must look for alternative ways of assessment that can provide the necessary data.

## **2 The properties of assessment**

The properties we want summative assessment to have are related not only to its validity and reliability as a measure of students' achievement but also to its impact on other components of the system, on teaching and learning and on students' motivation for learning. The effectiveness of assessment for a particular purpose cannot be judged in isolation. Any assessment is part of a whole system which includes: the way in which the data are communicated, reported and used; the preparation of teachers; the moderation of their judgements; the way in which evidence from assessment by teachers is combined with that from any external tests; the role of measures of the performance of students in the accountability of teachers and schools; and how the performance of students is monitored at local and national levels. Thus how results are used and the impact of this use has to be taken into account. This will affect the curriculum and pedagogy, as noted earlier, and it will also affect other part of the assessment system. For example, there is research evidence that:

- when school accountability is based on external summative assessment of its students, this impacts on the way that teachers conduct their own internal summative assessment and on how they use assessment formatively;
- how summative assessment for external uses is conducted influences teachers' own internal summative assessment;
- using student attainment data for accountability and system monitoring restricts the range of information that can be used for either purpose.

Assessment for any purpose should provide information with certain optimum properties: validity, desired impact, reliability and practicability. We consider each of these briefly before looking at how they apply to summative assessment.

### **2.1 Validity**

In this context validity means how well what is assessed corresponds with the behaviour or learning outcome that it is intended should be assessed, often referred to

as construct validity. Various types of validity have been identified, most relating to the type of evidence used in judging it (eg face, concurrent, content validity), but there is general agreement that these are contained within the overarching concept of construct validity. The important requirement is that the assessment concerns all aspects – and only those aspects - of students' achievement relevant to a particular purpose. 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.

## **2.2 Impact**

Here impact means the consequences of the assessment, often referred to as consequential validity (Messick, 1989). It concerns how appropriate the assessment information is for the uses that are made of it. As noted earlier, assessment generally has an impact on the curriculum and on pedagogy, so it is important that any potential adverse effects are minimised. Assessment can only serve its intended purpose effectively if this is the case.

## **2.3 Reliability**

The reliability of an assessment refers to the extent that the results can be said to be of acceptable consistency for a particular use. This may not be the case if, for instance, the results are influenced by who conducts the assessment or depend on the particular occasion or circumstances at a certain time. Thus reliability is often defined as, and measured by, the extent to which the assessment, if repeated, would give the same result. The extent to which high reliability is necessary depends on the purpose and use of an assessment. When assessment is used formatively, it involves only the students and the teachers. No judgement of grade or level is involved; only the judgement of how to help a student take the next steps in learning, so reliability is not an issue. Information is gathered frequently by the teachers who will be able to use feedback to the student to correct any mistaken judgement. However, high reliability *is* necessary when the results are used by others and when students are being compared or selected.

## **2.4 Practicability**

We use this term to mean that the resources required to provide an assessment are commensurate with the value of the information for 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 high degree of reliability is required. As well as the trade-off between these two properties (see below), there is a limit to the time and expertise that can be used in developing and operating, for example, a highly reliable external test. 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.

### **2.5 The relationship between reliability and validity and the concept of dependability**

The properties of reliability and validity are not independent of each other; attempts to increase one have the effect of decreasing the other<sup>3</sup>. To increase reliability the sources of variation are minimised, for instance by only using tasks that can be consistently marked or marked by machine. This reduces the opportunities to ask questions or set tasks that require more open responses and so to assess the outcomes that require such responses. The validity of the assessment in relation to such outcomes is thus reduced by the requirements of increasing reliability. On the other hand, if validity is increased by extending the range to include those outcomes that require more open responses, then reliability is likely to be reduced because more judgement is required in marking and grading such responses.

In recognition of the interaction of validity and reliability it is sometimes useful to refer to the combination of the two in the concept of dependability. In defining this concept, we give priority to validity; proposing that it means the extent to which the reliability is optimised while ensuring validity<sup>4</sup>.

## **3. Properties of summative assessment**

All educational assessment involves collecting evidence, making judgements and drawing inferences about learning from the evidence, reporting the findings in some way and using the findings for a given purpose. Evidence of what students can do can take a number of forms:

- written, as part of regular work or in response to special tasks or tests;
- oral, as part of regular work or in response to special tasks or tests;
- performance, as part of regular work or in response to special tasks or tests;
- artefacts, created as part of regular work or in response to special tasks or tests;
- portfolios of written work or artefacts, created as part of regular work or in response to special tasks or tests.

The selection of and combination of forms of evidence depends on relevance to the learning outcomes being assessed and the desired properties in relation to the purpose. For summative assessment of the outcomes identified in section 1.3 it appears that in various circumstances all forms of evidence above can be considered relevant. Selection then depends on the properties in relation to the purpose. Here it is useful to consider internal and external summative assessment separately.

### **3.1 Summative assessment for internal purposes**

Summative assessment for internal purposes is, in theory, under the control of the teacher, within the limits of the school or college policy on assessment, although it is subject to the influence of the requirements of external tests (see page 7). The

---

<sup>3</sup> See Chapter 7, The Reliability of Assessment by P. Black and D. Wiliam. In J. Gardner (ed) *Assessment and Learning* (2006) London: Sage

<sup>4</sup> See Working Paper 2, page 5

information is required for keeping records, evaluating progress of individual students, reporting to parents and students at regular intervals, passing information to other teachers on transfer from class to class or, in secondary schools, taking decisions about further subjects to study. It should concern all curriculum goals and provide evidence of the 'latest and best' performance of the students.

The evidence is most frequently gathered from regular work, from special tasks, teacher-made tests or externally developed tests. Although the evidence from these sources can be used to provide feedback to students and into teaching decisions, the main reason for collecting it is to check up on what students have learned from a series of lessons over a period of time.

It is not the evidence that is recorded but a judgement of it, which should be made in relation to the same criteria for all students. It most useful if there is a separate report for different areas of academic and non-academic achievement, communicated in a form that is understood by those who need the information.

### *3.1.1 Validity*

High validity is important for assessment for this purpose for three main reasons. First, it must provide a record of all the learning of each student. Learning that is not recorded becomes invisible. Second, it provides feedback to parents and the students themselves about learning and about what it is important to learn. Third, it may be used, as in some secondary schools, to make decisions about grouping of students, and about advice on their choices of subjects for continuing study. The fullest information is thus needed to inform these decisions. Where there is grading, evidence pertaining to a particular grade boundary might be crucial.

Validity will depend on how well the evidence and how it is interpreted reflects the learning being assessed. It is likely to be reduced if the assessment depends entirely or mainly on external tests, since these are less likely to provide information of higher validity for internal purposes than teachers' assessments, due to a poorer match of external tests to the learning goals over the relevant period. Similarly, validity may be low if the assessment depends mainly upon teacher-made tasks and tests designed specifically for assessment purposes, since these frequently focus on knowledge and information gained rather than the full range of goals of learning, including skills, attitudes and creative and critical thinking skills. The validity will also be infringed if the implemented curriculum is not consistent with the intended curriculum.

On the other hand, the validity is likely to be enhanced if different kinds of evidence are used to suit the range of learning outcomes to be reported. Validity is likely to be greater when teachers use evidence from regular activities, as these should cover the complete range of goals. This is evidence that can be collected and used for formative assessment and then reinterpreted against common criteria<sup>5</sup>. The validity will then depend on the nature of the evidence collected and on the quality of the interpretation.

---

<sup>5</sup> See Chapter 6 On the relationship between formative and summative purposes of assessment, in Gardner (Ed) (2006) *Assessment and Learning* London: Sage

Teachers have some freedom to decide how to achieve overall goals that are specified at certain points (eg the end of key stages in the English national curriculum). They also need freedom to decide what it is relevant to assess at intermediate points.

### *3.1.2 Impact*

For internal summative assessment this will depend not only on what information is gathered and how, as discussed above, but also on how frequently it is gathered. To reduce the unwanted consequences of internal summative assessment it should not be carried out too frequently, but only when summative reports are required.

Anticipating the summative assessment required for reports at the end of a term, or half-year, by assessing students in terms of levels or grades more frequently means that the feedback that students receive is predominantly judgemental, encouraging students to compare themselves with others. In such circumstances there is little attention to the formative use of assessment.

Unintended consequences are likely to be minimised by using evidence from regular activities, interpreted in terms of reporting criteria on a 'latest and best' basis and doing this when summary reports are required, not more frequently. Summative assessment should be carried out only at times when achievement needs to be summarised and progress evaluated; at other times teachers' assessment should be formative.

### *3.1.3 Reliability*

As noted earlier, where both reliability and validity are required to be optimal, there is a trade-off between them. In the case of summative assessment for internal purpose the trade-off can be in favour of validity, since no external decisions hang on the reported data. However threats to reliability need to be minimised. The nature of these threats depends on what evidence is used. If the evidence comes from tests or special tasks, then the reliability of the information will depend on the content, the marking and how the students feel about tests. If the evidence is derived from regular work, gathered over time by the teacher, the process avoids anxiety induced by tests but will depend on the interpretation of the evidence by the teacher. To increase reliability of their interpretation, some internal school procedures for teachers to moderate each others' judgments are necessary. This is likely to involve teachers of similar subjects or age groups meeting together to align their judgements of particular pieces of work, again representing the 'best evidence' on which the record or report is to be made. Moderation should include responses to any teacher-made tests, for it is a mistake to assume that because they have greater formality tests are automatically more accurate than judgements based on regular work.

### *3.1.4 Practicability*

Assessment for this purpose is essentially carried out by the teacher but in order to ensure adequate reliability, as noted above, some in-school moderation is necessary. Some tentative estimates of the teacher and student time involved in various assessment-related activities are given in the Appendix. This shows that from Y1 (5 to 6 year olds) to Y4 (8 to 9 year olds) teachers spend the equivalent of 3 to 4 hours per week on on-going assessment, marking, moderation, report writing and parents' evenings. This rises to almost 6 hours per week in Y5 and Y6. Pupils' time spent on assessment-related activities amounts to about 30 minutes per week in Years 1 – 4 and

several hours per week in Y5 and Y6 due to teachers giving more tests and preparing for end-of-Key-Stage tests.

At Key Stage 3 it appears that little time is spent on moderation, teachers depending for internal summative assessment on tests. Some teachers are spending as much as 90 hours a year per class on assessment-related activities, about a third of this being directly related to testing and two-thirds relating to assessment of class work and general marking. Pupils are spending 18% of their lesson time on assessment, which is about 40 minutes for each subject or nearly 6 hours per week across all subjects. In Key stage 4 this rises to an hour per week per subject.

At Key Stage 4 and for post 16 pupils the costs of assessment shift from the indirect cost of teachers' time to the direct costs of external examination entrance fees and administration and invigilation costs. Teachers are spending about an hour per week per class on marking and about the same on coursework assessment. The cost in students' time is about three and a half weeks of lesson time per year in each subject.

### **3.2 Summative assessment for external purposes**

This purpose requires a public statement of what a student has achieved for use by those outside as well as inside the school. The process requires more rigour than in the case of summative assessment for internal uses, since the result may be used for selection or comparison among students from different schools and for evaluating teaching. In these uses there is an assumption that the summary grade, level or standard achieved means the same for all students no matter who makes the judgment. That is, reliability must be high enough to give confidence in these outcomes of assessment. However, the requirements of high validity are equally as strong as for internal summative assessment. Thus the emphasis has to be on increasing reliability in the interpretation of the evidence.

To serve the intended purposes, the information should:

- provide a summary of what the student has achieved at a certain time, based on the 'best evidence';
- reflect the full range of relevant goals;
- be based on evidence judged by the same criteria for all students in all schools, with quality assured by moderation;
- be reported in terms of levels or grades meaningful to users.

#### *3.2.1 Validity*

High validity is essential, first because it sends a strong message about what learning is valued and second, because of the high stakes attached to summative assessment for external purposes. In most cases of assessment for external purposes the stakes are already high for the individual student. Adding high stakes for the teacher, by using the results for evaluation of teaching, means that the assessment is bound to have a strong influence on teaching and learning, in whatever way it is carried out. This is taken further in the discussion of impact. The point to make here is that the assessment will not give the information that is needed by those receiving and using the information if it does not reflect the full range of goals but only a narrow range of outcomes that are more reliably assessed.

Most users want to see evidence of both academic and non-academic achievement. The evidence collected at the project's seminars show that students and parents want all students' achievements, including what is learned outside the classroom, to be reported and credited; both employers and higher education admission tutors want to be able to identify students who are independent learners, who show initiative and perseverance and have learned how to learn. Reporting such outcomes when students leave school is not enough; the progress towards them needs to be monitored throughout school. Consequently such outcomes must be included in valid summative assessment.

How well the assessment reflects the range of achievements that it is intended should be included will depend not only on what evidence is collected but on how it is collected. In relation to 'how', while some learning outcomes among those suggested in 1.3 can be assessed by machine-markable tests, most require evidence of different kinds, some from open-ended tasks or tests and others from evidence gathered over time. Validity is threatened when concern for reliability leads to a reduction in how well what is assessed reflects the range of achievement that is intended and required by users of the assessment outcome. Examples are the science practical examinations taking the form of routine procedures making little cognitive demand on students, but easily scored. This threat can be minimised by using a wider range of evidence including teachers' judgements across a range of activities. Where teachers' judgements are used for this purpose, a greater degree of quality assurance is needed than for assessment for internal purposes, involving professional development and inter-school moderation.

Use of evidence gathered and judged by teachers can improve the match between the range of intended learning and the information provided in the assessment since, as part of their regular work, teachers can build up a picture of students' attainment across the full ranges of activities and goals. This gives a broader and fuller account of achievement than can be obtained through tests, which can only include a restricted range of items. Freedom from test anxiety means that the assessment is a more valid indication of students' achievement.

In some circumstances it may be desirable for teachers' judgements to be supplemented by information from externally devised tasks or tests. The reason may be that these tasks present the best way of providing evidence of certain skills or understanding, or that the use of the assessment data requires a greater degree of uniformity in how the assessment is conducted than is found in teachers' assessment data. A well designed set of assessment tasks available for teachers to use has several benefits. Such tasks can exemplify for teachers the situations in which skills and understanding are used and thus guide them in developing their own embedded assessment tasks. They are also of particular benefit to newly qualified teachers and those who need to build their confidence in their ability to assess students.

### *3.2.2 Impact*

Teaching will inevitably be focused on what is assessed. The impact of summative assessment on student learning experiences can be positive if the assessment covers the full range of intended goals, when the assessment criteria often help to clarify the meaning of the goals. However the impact on learning experiences can be restrictive if there is a mismatch between the intended curriculum and the scope of the

assessment. The consequences are likely to be more severe when results are used for accountability of teachers and schools. When the information is gathered from external tests and examinations this puts pressure on teachers, which is known to lead to coaching in how to pass tests, multiple practice tests and shallow learning<sup>6</sup>. Validity is infringed because results no longer signify learning achieved but rather the ability to answer test or exam questions. Other known consequences are the demotivation of lower achieving students and, for all students, a view of learning as product rather than process.

Some of these consequences can equally follow from using teachers' judgements for summative assessment if the results are used for accountability. In such circumstances, moderation procedures can become over-elaborate and constrain teachers to collecting evidence using 'simple and safe' methods rather than more educationally valuable ones. Then there is likely to be a tendency to focus on a narrow interpretation of criteria and on *performance* rather than *learning*. This is best avoided by basing accountability on a wider range of information about the context and outcomes students' learning and on school self-evaluation. (See section 5.2)

### 3.2.3 Reliability

Summative assessment for external purposes is conducted to inform decisions. These may affect individual students, as in selection for secondary schools, or entry to further or higher education courses. Even when summative assessment results do not have serious implications for individual students (as in national tests) they can acquire high stakes when the results are used for evaluating teaching or ranking schools. In either case, the accuracy of the information is of the essence.

When the assessment is carried out using tests or examinations, striving for reliability can infringe construct validity since in the development of instruments there is a preference for items relating to outcomes that are most reliably assessed. These most often include factual knowledge, where answers can be marked unequivocally, and exclude what is more difficult to judge, such as application of knowledge, critical reasoning and affective outcomes. We have argued a strong case for including these outcomes. Thus it is necessary to find a way of assessment them with acceptable reliability. Assessment by teachers has the potential for providing information about a wide range of cognitive and affective outcomes but the reliability of teachers' judgements is widely held to be low and there is research evidence of bias<sup>7</sup>. However, the research also shows that when criteria are well specified (and understood) the teachers are able to make judgments of acceptable reliability. The moderation procedures that are required for quality assurance can be conducted in a way that provides quality enhancement of teaching and learning as noted earlier in the context of summative assessment for internal school use.

---

<sup>6</sup> There is a considerable body of research to support these statements. A summary of the most dependable evidence is given in the ARG booklet entitled *Testing, Motivation and Learning*, which can be downloaded from the ARG website (see note 1).

<sup>7</sup> A brief summary of this evidence is given in ASF Working Paper 2 (available from the ARG website) and in Harlen, W. (2005) *Trusting teachers' judgment: research evidence of the reliability and validity of teachers' assessment used for summative purposes*. *Research Papers in Education* (2005) 20 (3), 247-270

Discussion of what is ‘acceptable’ reliability is not well informed by relevant data. It is generally assumed that the information provided by tests and examinations is necessarily more reliable than that derived from teachers’ judgments. However, the estimates of misclassification by standardised tests in England<sup>8</sup> are quite alarming and there are good reasons to question whether assumptions of high reliability are justified. The high incidence of successful appeals reflects unreliable marking; but this is not the main area for concern. Because the number of items and time for answering is limited, what is tested is only a small sample of the work covered and a different selection of items could easily produce a different result for particular students.

### 3.2.4 *Practicability*

Estimating costs of assessment is notoriously difficult; the uncertainty of concepts and the complexity of variables mean that all reports have to be treated with great caution. The appendix provide some very tentative estimates of the time and direct costs of summative assessment, including teacher-made and external tests and examinations in primary and secondary schools in England, where national testing takes place in Y2 (7 year olds), Y6 (11 year olds) and Y9 (14 year olds) and examinations are taken at the end of key stage 4 (aged 16) and in the following two years (AS and A2 level).

#### Key stages 1 and 2 (England)

The estimate for teachers’ time spent on all assessment-related tasks, including teachers’ assessment, moderation, report writing, parents; evenings and national testing where appropriate, varies from 130 hours in Y1 to 160 hours in Y6. Teachers’ time spent specifically on testing (both internal and external) varies from zero in Y1 to 20 hours per year in Y6. The actual time spent on national testing is much smaller than the time spent on regular tests given by the teacher. Other evidence would suggest that this is the result of teachers giving students practice tests and of assuming that tests are necessary, in preference to their own judgements. The amount of testing conducted by teachers would be expected to fall if teachers’ judgments were more widely used and trusted, with teachers then being able to spend up to 12% (based on a working year of 1265 hours) of their time in other ways.

A move to making greater use of teachers’ judgements in external summative assessment means more time for moderation. To increase this to, say, the equivalent of half a day every 3 weeks (one third of PPA time) requires a total of about 50 hours per year. Since at least half of this is already included in the time spent on assessment, the increase would easily be absorbed into the 12% of time saved by reduction of testing.

Students’ time spent on summative assessment activities varies from zero in Y1 to 9% in Y6. There is a considerable leap in time on assessment from Y4 to Y5, where 7% of time is being spent. The national tests themselves (including revision) take up 6% of time in Y6. The estimates in Appendix A suggest that if the time spent on assessment was reduced to that needed for regular formative and summative purposes, this would be the equivalent of adding two weeks of learning time to the school year.

---

<sup>8</sup> See, for example, Chapter 7: The Reliability of Assessments by Paul Black and Dylan Wiliam in Gardner (Ed) *Assessment and Learning* London: Sage, 2006.

### Key Stage 3 (England)

At Key stage 3, the figures are based on a time spent in various activities by science teachers, which can perhaps with caution be generalised to other subject areas. The figures show that teachers spend 90 to 100 hours per class per year on assessment, the majority on their own regular assessments, including marking. National tests and other tests such as end of module tests take about 30 to 35 hours and other assessments about 54 hours (with some further time being taken in report writing and parents' evenings). As noted earlier, currently little time is spent on moderation in KS3. Fewer tests and more assessment by teachers would increase the need for moderation. However, the time saved in testing more than compensates for the 3% of time necessary to provide half a day every three weeks for moderation activities.

The student time taken by assessment and external tests is about 18% of learning time. If half of this were saved by reduction in external testing, a further three weeks of learning time could be added to the school year.

### Key Stage 4 and AS/A2 level

Since modularisation, the time spent on assessment has increased and external end-of-module tests have replaced a good deal of the teacher-made tests and mock examinations used before modularisation. Teachers are reported as spending about 8% of their time on assessment related activities in both Y12 and Y13. Half of this is concerned with tests, examinations and coursework assessment. . How this time is spent varies from subject to subject. For example, in science at A level half of this is spent on examinations and half on coursework. This low figure, compared with Key Stage 3, reflects the dependence on external end of module tests and examinations. Students are spending about 28 hours per year per subject, or roughly an hour a week, on tests, examinations and coursework assessment.

How these costs would change with greater use of teachers' assessment is not known. But it seems clear that at Key Stage 4 and for post 16 pupils the costs of assessment shifts from the indirect cost of teachers' time to the direct costs of external examination entrance fees and administration and invigilation costs. Savings here would provide the funding necessary for teachers to spend more time in moderating their judgements, since the average 11-18 secondary school spends the equivalent of the salary of three teachers on examination costs. While not all of these costs of external awards would be saved by teacher assessment, a considerable reduction is certainly possible. Further, some of the 60 to 70 hours per year that teachers are already spending on coursework assessment and tests and examination could be used for quality assurance of their judgements. The saving in students' time would be about three and a half weeks of lesson time per year in each subject.

## **4. The properties of summative assessment by teachers**

To inform decisions about the role that assessment by teachers can take in summative assessment Table 1 brings together the properties of summative assessment by teachers in comparison with those of tests.

Table 1

Property	Summative assessment by teachers	Summative assessment by tests
Validity	Potential for the full range of goals reflecting the whole of the curriculum. Freedom from test-anxiety and from practice in test-taking means that students can show what they can do in normal conditions. Validity depends on opportunities provided in teaching.	A sample only of the full range of goals and a sample only of those goals that are assessed. Ensures that all students are judged on the same items and tasks. Schools ensure that all students are taught what is required by the awarding bodies' specifications.
Impact	Reflects and reinforces what is taught; can use evidence from formative assessment. Provides opportunities for students self-assessment. Public confidence in results may be low in comparison with results of examinations.	Leads to coaching what is tested, teaching test-taking skills, and a summative ethos in classroom assessment. Provides students with goals and public with 'gold standard' by which to judge students and schools.
Reliability	Perceived as being unreliable and biased. Judgments require moderation appropriate to the use of the data. With appropriate training can reach levels similar to those of tests.	Subject to measurement and marking errors such that up to 30% of students may be misclassified <sup>9</sup> Some external tasks or tests are needed to ensure external confidence in comparability across schools.
Practicability	Increase in teacher workload due to additional responsibilities. Training and moderation essential. Less external testing likely to mean less use of commercial tests in preparation for them. Teacher time released from preparing and marking tests. Students' learning time increased.	Takes large proportions of teaching and learning time. High cost to schools for entrance fees and for administering external examinations. Separates roles of assessor and teacher.

Since teachers already have a major role in summative assessment for internal purposes, the major area of change relates to assessment for external purposes. However, the steps needed to ensure that this has the required properties discussed in section 3 will equally serve to improve practice of summative assessment for internal purposes.

### **5. The relationship between summative assessment of students and assessment for other purposes**

The interaction between components of the assessment system as a whole was noted at the start of section 2. There is research evidence showing that how assessment for

<sup>9</sup> Wiliam, D (2001) Reliability, validity, and all that jazz. *Education 3-13*, 29 (3) 17-21

one purpose is conducted can affect not only the curriculum and pedagogy but also other parts of the assessment system. In the present context we are concerned particularly with the relationships between:

- summative assessment and formative assessment
- summative assessment and accountability
- summative assessment and system monitoring

### **5.1 Summative assessment and formative assessment**

Among the advantages of using assessment by teachers for summative assessment are:

- avoiding the single, end of course test or examination, with all its negative impacts on students, teaching and the curriculum;
- greater freedom for teachers to pursue learning goals in ways best suited to their students, rather than being constrained by what is perceived as necessary in order for students to pass tests;
- using evidence about students' on-going achievements formatively, to help learning, as well as for summative purposes;
- facilitating a more open and collaborative approach to summative assessment in which students can share in the process through self-assessment and derive a sense of progress towards 'learning goals' as distinct from 'performance goals'.

In WP1 we pointed out that evidence gathered about the progress of a student towards the goal of a particular piece of work (which may be chosen to suit the individual needs of the student) is used formatively when it is used to inform feedback about the next steps the student needs to take. The feedback to students should indicate these next steps but should not, if it is to be really useful to the students, be in terms of a judgement about what level or grade has been achieved. For summative assessment, the information from work over a period of time (including ephemeral evidence) is brought together and judged in relation to grade or level criteria that are the same for all students. Teachers should be able to relate student-specific goals to the goals for all students. The evidence from specific lessons, used in helping students to achieve lesson goals, has to be brought together and reviewed against the broader criteria that define reporting levels or grades. This involves finding the 'best fit' between the evidence gathered about each student and one of the reporting levels<sup>10</sup>, giving preference to the 'best evidence', which may not necessarily be the 'latest'.

### **5.2 Summative assessment and accountability**

Accountability in this context implies being answerable to oneself or to another person or body for the use of public resources. For teachers, accountability operates at a number of levels: to themselves, at the individual level, as part of self-evaluation; to the school and its parents and governors, as part of its internal evaluation; to the local authority, as part of external school evaluation; and so on to the national level. The effect of accountability at each of these levels depends on:

---

<sup>10</sup> See reference in note 6

- the type of information that is taken into account (eg varying from student achievement data only, to a range of relevant information about learning processes, contexts and outcomes);
- the criteria used in judging effectiveness (eg in relation to student achievement varying from whether externally prescribed targets in a narrow range of knowledge and skills have been met to whether achievement in a range of academic and non-academic meets expectations);
- the action that follows the judgement of effectiveness (eg varying from sanctions or rewards, to supportive action to correct areas of weakness).

When the information is derived from summative assessment carried out for other purposes, there is the danger that it is not well matched to the purpose of accountability. For instance, there are multiple disadvantages when external test or examination results are used and a school is held accountable for whether a specified target percentage of students reach a certain level. The results are unlikely to reflect the full range of educational outcomes which a school strives for and for which it should be held accountable. Further, to reach the target, attention is focused on those students who are close to the required level, with less attention to those who are either too far below or are already above the target level. A third point is the focus on the narrow requirements of passing the test or examination.

Thus framing accountability in terms of targets for student achievement, or position in a league table of schools based on test and examination results, distorts the actions of those held accountable in ways that are not intended and are not in the best interests of students. It means that neither is there genuine information about actions for which teachers and schools should be held responsible nor a positive impact that enables performance to be improved.

For a more positive impact, accountability is best based on information about a range of student achievements and learning activities, judged by reference to the context and circumstances of the school and used positively to improve students' opportunities for learning. It follows from these arguments that the information used in accountability should include information about the curriculum and teaching methods and relevant aspects of students' backgrounds and of their learning histories. Some good examples exist in various school self-evaluation guidelines<sup>11</sup>.

### **5.3 Summative assessment and system monitoring**

Monitoring implies the regular collection of information in order to detect changes over time. In the context of education it refers to changes in levels of student achievement and is usually associated with whether 'standards' are rising, falling or remaining steady. Although the evidence used includes student achievement, the purpose is to inform policy and practice decisions, not to make judgements or decisions about individual students. Monitoring can concern a range of aspects of practice but generally requires equivalent evidence to be gathered about different cohorts of students at a particular stage or age, so that, say, performance of 13 year

---

<sup>11</sup> For example: *How Good is Our School* in Scotland, in England the emphasis on schools self-evaluation in 'A New Relationship with Schools' (DfES and Ofsted, 2004) and in Wales *Guidance on the Inspection of Primary and Nursery Schools* (September 2004) and *Guidance on the Inspection of Secondary Schools* (September 2004) (Estyn).

olds at one time can be compared with performance of 13 year olds at a later time. Since there are many differences between successive cohorts of students of the same age, comparisons for small groups have very low reliability. Monitoring at the school level is best undertaken with the context of self-evaluation, where other information needed to interpret student assessment data is also collected. Even at the system level, which is our concern here, a change from one year to the next is unlikely to be meaningful; trends over time provide more useful information.

The value of system monitoring depends on the range of information that is collected. The rather obvious approach of collecting test results from national tests, taken by every students in a cohort, reduces the information to the small sample of the subject domain that any one students can be reasonably expected to provide. This course of action reflects a fundamental confusion between the sample of items that a student can take, with the sample that is adequate for monitoring performance across the domain. The validity is further reduced by basing the information only on the results of tests that are designed to give reliable information about individuals and thus focusing on aspects that can be scored most easily and objectively. For valid monitoring a wider range of evidence is needed, derived from observation of skills in action as well as assessment of products.

The economical advantage of collecting achievement data already available, as in using national tests for identifying national trends, must be judged against the extent to which it provides useful and relevant information. Similarly the more costly process of establishing and running surveys covering a wide range of educational outcomes (as in the APU surveys in England, Wales and Northern Ireland in the 1980s and the on-going Scottish Surveys of Achievement) has to be judged against providing more detailed feedback that can be useful at the policy level, but also directly to practitioners. Separating monitoring from the performance of individual students would obviate the need for central collection of student assessment data. In turn, this would set student summative assessment free from the high stakes that restricts what is assessed and what is taught to what can be tested.

## **6. Implications for the role of assessment by teachers in a summative assessment system**

Here we bring together the main points from the earlier discussion that focus on the use of assessment by teachers. One of the implications of the arguments set out in Working Paper 1 was that:

A comprehensive summative assessment system must be capable of providing information, based on dependable judgments, about how much and how well students have developed a wide range of cognitive competencies and affective outcomes.

The further discussion in this paper has led to the conclusion that assessment by teachers has a key role in such a comprehensive and dependable system. The main reasons for this relate to validity and to the impact on students, teachers and teaching, including the practice of using assessment formatively.

Using teachers' assessment enhances the validity of summative assessment because it enables a wider range of learning outcomes to be included than can be done using formal tests and examinations. To prepare students adequately for adapting to changing occupational requirements and communication modes, the goals of learning must include higher level skills, the development of understanding and the understanding of learning. Teachers' assessment has the potential to provide information about these processes and outcomes of learning and so ensure a better match between the intended curriculum and the scope of the assessment.

In terms of impact, there is clear research evidence that testing has a narrowing impact on the curriculum, on teaching methods and on students' motivation for learning. It also leads to more summative assessment than is really necessary, through encouraging practice tests and increased use of commercial and teacher-made tests. A serious consequence of this is to reduce the use of assessment to help learning. The known value of formative assessment in raising standards and reducing the gap between the higher and lower attaining students, means that opportunities to improve students' learning are being lost. Whilst using teachers' assessment for summative assessment will not necessarily mean more use of formative assessment (particularly if the results are used for high stakes judgments of teachers and schools), a combination of help to implement formative assessment and training in applying criteria to students' regular work can substantially reduce the negative impact (see the case studies in Working Paper 4).

However, any system that makes use of teachers' assessment for summative purposes has to address the concerns about dependability that are prevalent among users of summative assessment. It is also recognised that the assessment that is currently undertaken for internal school purposes often needs to be improved. The project's discussions with students, parents, employers, higher education<sup>12</sup> and teachers themselves, underlined general concern about:

- 'fairness' being compromised by teachers favouring certain students or by students having help from others with their work;
- whether all teachers can apply the same standards in making their judgements;
- the extra work load for teachers that is assumed to be involved;
- teachers' assessment being regarded as inferior to external tests.

The discussion in Section 3 indicates, in general terms, ways in which these concerns can be addressed, while accounts of procedures being implemented or proposed in specific circumstances are outlined in Working Paper 4. The main points, which echo the implications set out in Working paper 1, are:

- Summative assessment, by teachers or other means, should be designed so that it provides information for specific purposes and carried out only at times when achievement needs to be summarised and progress evaluated; at other times teachers' assessment should be formative.
- Both pre-service and in-service professional development should specifically address: the development of teachers' understanding and skills of assessment for different purposes; the potential bias in teachers' assessment; and help teachers to minimise the negative impact of assessment on students.

---

<sup>12</sup> See Report of ASF Seminar 5 on the ARG website

- Attention and resources must be given to creating developmental criteria, based on evidence of how students learn, which indicate a progression in learning related to particular goals and can be applied to a range of relevant learning activities. Procedures to enhance teachers' own summative assessment should reflect this priority
- Robust and permanent procedures for quality assurance and quality control of teachers' judgments are needed to ensure that teachers' summative assessment provides valid and reliable accounts of student learning.
- Assessment procedures need to be transparent and judgements supported by evidence so that all users know how results were obtained.
- Teacher should have access to optional tasks assessing skills and understanding, which they can use to assist them in making judgments across the full range of learning goals.
- Summative assessment systems making greater use of teachers' assessment should be designed to enable teachers to use their time effectively, minimising the burden on teachers and students.
- Summative assessment must be in harmony with the procedures of formative assessment, supporting the use of assessment by teachers and students to help learning.

Further, to avoid the negative impact of the 'high stakes' that follow from using summative assessment for evaluating teachers and schools:

- Systems of school accountability should not rely solely, or even mainly, on the data derived from summative assessment; such data should be reported, and interpreted, in the context of the broad set of indicators of school effectiveness.
- The monitoring of standards of students achievement should be derived from a wider base of evidence than test results from individual students. Teachers' assessment has a place in a system in which a wide range of evidence is collected for small samples of students.

### **References**

- Black, P., McCormick, R., James, M. and Pedder, D. (2006) Assessment for learning and learning how to learn: a theoretical inquiry. *Research Papers in Education* in press
- Black, P.J. and Wiliam, D (1998) Assessment and Classroom Learning. *Assessment in Education*. 5 (1) 7-71
- Bransford, J. D., Brown, A.L. and Cocking, R. R. (Eds) (1999) *How People Learn: Mind Brain, Experience and School*. Washington: National Academy Press
- Crooks, T.J. (1988) The impact of classroom evaluation practices on students. *Review of Educational Research*, 58, 438-481
- Harlen, W. and Deakin Crick, R. (2003) Testing and motivation for learning. *Assessment in Education*, 10(2) 169-207
- Messick, S. (1989) Validity. In (Ed) R. Linn *Educational Measurement 3<sup>rd</sup> Edition*. New York: Macmillan and American Council on Education 13-1-3
- OECD (1999) *Measuring Students Knowledge and Skills, A New Framework for Assessment*. Paris: OECD.
- White, R.T. (1988) *Learning Science*. Oxford: Blackwell
- Wiske, M. S. (Ed) (1998) *Teaching for Understanding*. San Francisco: Jossey-Bass

### **Estimated time and direct costs of summative assessment, including teacher-made and external tests and examinations in primary and secondary schools in England**

#### **Sources of information**

It was beyond the means and scope of the project to undertake a survey of the costs of tests and examination. Such a survey was undertaken by PriceWaterhouseCoopers (PWC)<sup>13</sup> for the QCA, between September and November 2003, the report noting that ‘no such exercise has previously been undertaken in this country, or indeed overseas’ (p 6). Whilst spelling out their mapping methodology, however, they provide only overall cost figures in the published report. A second source of information was a report for the Royal Society from the Centre for Science Education of Sheffield Hallam<sup>14</sup> of a survey of the cost of testing and assessment in science published in March 2003. This survey has the advantage of providing some detail about how time is spent on all assessment activities and not just on statutory tests and examinations.

However, there are doubts about the dependability of the results of both these surveys, since it is almost impossible to separate assessment activities from teachers’ other activities in a consistent and meaningful way and, in the case of the Sheffield Hallam report, to extrapolate to the whole curriculum. *Thus these results can only be taken as broad indications of the order of magnitude of time and costs involved.*

Reference has also been made to a survey of secondary schools by the Secondary Heads Association (SHA). Finally, actual figures from individual schools have been used to set beside the general picture from these surveys.

#### **Overall costs: time and direct costs**

Attention is most often given to the direct costs to individual schools and to the system as a whole. But just as important is the cost in terms of teaching and learning time. Thus we focus here on these time costs, at the school level. Teacher time costs can be turned into system costs by scaling up and multiplying by the average costs of employing teachers. When this is done, the PWC report arrives at the figure of £240m for time involved in the process of administering tests and examinations in primary and secondary schools. This does not include other assessment activities, such as marking, writing reports, parents’ evenings, etc., which are included here in the estimates of the total time spent in assessment and testing at various stages.

---

<sup>13</sup> *Financial Modelling of the English Exams System 2003-4*, report from PriceWaterhouseCoopers (PWC) for the QCA (2004)

<sup>14</sup> *The Cost of Assessment*. Centre for Science Education, Sheffield Hallam University 2003

Direct costs are those of providing, administering, invigilating, marking, and reporting tests and examinations. These costs are borne variously by QCA, the Awarding bodies, schools and colleges. According to the PWC report they total £370m for all key stage tests, GCSEs, AS and A-levels, BTEC, GNVQ, AEs and FSMQs. The PWC estimate of total of time and direct costs for tests and examinations in England in 2003 is £610m.

The SHA survey includes secondary schools and sixth form colleges only in England and Wales and excludes key stage 3 tests. It is, therefore, very difficult to compare the overall result with that from the PWC report. The SHA survey concluded that the total cost in sixth form colleges and schools (but excluding general further education colleges) in England and Wales of external examination fees, administration and invigilation is £380 million.

### **Time spent on assessment-related activities**

#### **Key Stage 1: Class teachers' time used in assessment activities**

The summary, in Table 1, of hours spent on various assessment activities is based on information obtained from one school

*Table 1 Key Stage 1: Teachers' time (in hours per year)*

	Foundation	Y1	Y2
Teachers' assessment (observation, discussion, marking, preparation and use of any special tasks)	72	45	53
National testing (including marking)	n/a	n/a	20
Moderation	40	40	40
Report writing	30	30	30
Parents' evenings	15	15	15
Total	157	130	158

Assuming a 33 hour week this is equivalent to 4.7 weeks of teacher time in the foundation stage, 3.9 weeks in Y1 and 4.7 weeks in Year 2. The figures in Table 1 compare reasonably well with the extrapolation from the Sheffield Hallam report on science. The estimates in Tables 2 and 3 for Key Stages 2 and 3 are based on these extrapolations.

#### **Key Stage 2 Class teachers' time used in assessment activities**

In this case time taken for end of unit tests, including use of commercial tests and regular teacher-made tests, has been separated from other internal teachers' assessment. This provides a separate estimate of time spent on testing and special tasks and on regular assessment activities such as marking, observation and discussion.

*Table 2 Key Stage 2: Teachers' time (in hours per year)*

	Y3	Y4	Y5	Y6
Teachers' assessment (including observation, discussion, marking)	105	105	157	157
Internal testing and preparation and use of any special tasks or	96	96	96	150

commercial tests				
National testing	n/a	n/a	n/a	15
Moderation	25	25	25	30
Report writing	20	20	20	20
Parents' evenings	15	15	15	15
Total	261	261	313	387

These totals are equivalent to 7.9, 7.9, 9.5 and 11.7 weeks of teacher time, for Y3, Y4, Y5 and Y6, respectively.

**Key Stage 3: Subject teachers' time (in hours per year)**

Assuming that a science teacher's experience is typical of other secondary subject teachers' experience, the figures given by the Sheffield Hallam report can be used to estimate the average time taken in Years 7, 8, and 9. This average conceals an enormous variation among schools in certain respects. For instance no time at all was reported as being spent on moderation in some schools and this was the case for most for Y7 and Y8. In Year 9, some schools spent 12 hours per teacher per class and some none. Time spent on internal testing also varied widely, a Technology College administering 12 tests per year, with a comprehensive school giving three. The total time on these tests was estimated at 114 hours per class for the former and 43.5 hours per class for the latter. The average used in the overall estimate is 54 hours per class. The figures in Table 3 need to be read with these variations in mind.

*Table 3 Teachers' time: hours per class and total based on two classes per teacher per year group*

Activity	Y7	Y8	Y9	Total
Dealing with KS2 data	2			4
Provision of data to LEA	2.5			5
Module and other regular tests	28	28	20	152
National tests (mock and statutory, including organisation)	n/a	n/a	15	30
Other teacher assessment (practical tasks and marking)	54	54	54	336
Moderation	-	-	6	12
Report writing	4	4	4	24
Parents' evenings	3.3	3.3	3.3	19.8
Total	93.8	89.3	102.3	582.8

The total for a teacher is based on the assumption that a teacher would teach 18 out of 33 hours per week, the equivalent of two classes in each KS3 year.

### Time spent by teachers on summative tests in KS 1, 2, and 3

*Table 4 Total assessment time per year per class spent by teachers with and without summative tests.*

	Time in current regime (hours per class)	Time without summative tests (hours per class)	Time spent on external summative tests (hours per class)
Y1	130	130	0
Y2	158	138	20
Y3	261	165	96
Y4	261	165	96
Y5	313	217	96
Y6	387	222	165
Y7 (science only)	94	66	28
Y8 (science only)	89	61	28
Y9 (science only)	103	67	35

Thus the primary teachers up to 165 hours, and for secondary teachers at KS3 up to 35 hours per year per class, of teachers' time could be used in other ways if external tests were replaced by teacher assessment.

### Pupil time, Key Stages 1, 2, and 3

The figures in Table 5 are based on those given in the Sheffield Hallam report for the time spent by pupils in various activities undertaken specifically for summative assessment.

*Table 5 Pupils' time on summative assessment activities*

	Time in hours	Time as % of learning time
Y1 all subjects	-	0
Y2 all subjects	27.0	3 (of total 950 hours)
Y3 all subjects	13.5	1 (of total 950 hours)
Y4 all subjects	13.5	1 (of total 950 hours)
Y5 all subjects	66.0	7 (of total 950 hours)
Y6 all subjects	84.4	9 (of total 950 hours)
Y7 Science only	20.0	18 (of total 114 hours)
Y8 Science only	20.0	18 (of total 114 hours)
Y9 Science only	20.0	18 (of total 114 hours)

The large increase in time spent in assessment in Y5 and Y6 reflects a considerable increase in assessment in preparation for the national tests. Assuming that the time required for regular formative and summative assessment increases to about 20 hours in Y5 and Y6, then about 46 hours or 9 days in Y5 and 64 hours or 13 days in Y6 could be used in other ways instead of practising and taking tests. In Y 7-9, assuming about half of the assessment time is used on giving tests, then over three weeks of science lessons could be used in other ways.

## **Key stage 4: Pupil and teachers' time**

Science is a subject that can be studied in various courses leading to different awards at GCSE (single, double, 3 separate science, double award vocational). For double award science, teachers are estimated to spend 32% of their lesson time on assessment-related activities. The Sheffield Hallam report suggests that the testing and examinations element of this is 10% of lesson time.

KS4 pupils are spending about 30 hours a year on assessment-related activities in science, 17 hours of which are spent on test and examinations. This is almost 4 weeks of lessons time that could be spent in other ways.

## **AS/A2 level**

Teachers are reported as spending about 30% of lesson time on assessment related activities in both Y12 and Y13. 7% is concerned with tests and examinations and 9% with coursework assessment. Since modularisation the time spent on assessment has increased and external end of module tests have replaced a good deal of the teacher-made tests and mock examinations used before modularisation.

Pupils are spending about 3.5 weeks in Y12 and Y13 on assessment activities, about half on examinations and half on assessed coursework. As at KS4, this could be roughly halved if tests and examinations were replaced by teacher assessment of course work.

## **Direct costs**

The direct costs to schools of key stage national tests is small. There are direct costs for purchase of commercial tests and of packages of tests and related data handling, for example from CEM which cost about £3 per pupil.

However, the cost of external examinations for secondary schools are considerable. The three most expensive elements are:

- Examination fees
- Administration time (carried out by support staff from September 2003)
- Invigilation (carried out by support staff from September 2005)

The SHA survey of secondary schools and sixth form colleges reported that the average total cost to college and school budgets for examination fees, administration and invigilation is around £300,000 in a 1500-student college, £130,000 in 11-18 schools and £67,000 in 11-16 schools.