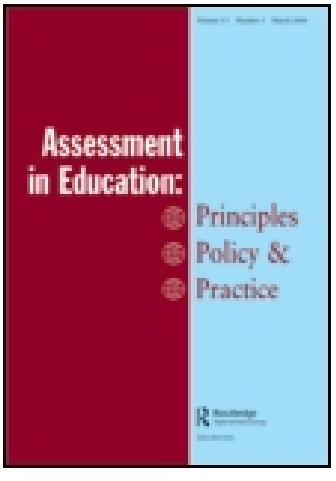

Assessment in Education: Principles, Policy & Practice
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caie20

Can teachers' summative assessments produce dependable results and also enhance classroom learning?

Paul Black [a], Christine Harrison [a], Jeremy Hodgen [a], Bethan Marshall [a] & Natasha Serret [a]
[a] Department of Education and Professional Studies, King's College, London, UK
Published online: 07 Nov 2011.

To cite this article: Paul Black, Christine Harrison, Jeremy Hodgen, Bethan Marshall & Natasha Serret (2011) Can teachers' summative assessments produce dependable results and also enhance classroom learning?, Assessment in Education: Principles, Policy & Practice, 18:4, 451-469, DOI: 10.1080/0969594X.2011.557020

To link to this article: http://dx.doi.org/10.1080/0969594X.2011.557020

Assessment in Education: Principles, Policy & Practice
Vol. 18, No. 4, November 2011, 451–469

Routledge
Taylor & Francis Group

Can teachers' summative assessments produce dependable results and also enhance classroom learning?

Paul Black*, Christine Harrison, Jeremy Hodgen, Bethan Marshall and Natasha Serret

Department of Education and Professional Studies, King's College, London, UK

Summative assessments that are integrated within the daily pedagogy of teachers are problematic. Some argue that they cannot both be helpful to pedagogy and yield results that are comparable across and between schools. Others claim that there is enough evidence to show that these targets can be achieved. The project described in this paper explored how teachers might enhance their competence in summative assessment in ways which might also have a positive effect on their teaching and learning. A strategy was developed based on five key features of summative assessment practices. The findings, from a longitudinal study with 18 teachers, are based on the teachers' opinions, both about the quality of the results which they achieved, and about the positive impacts on the involvement of pupils, on collaboration between teachers, and on interaction with parents. The project involved teachers of English and mathematics in three schools, working with the authors, over two-and-a-half years.

Keywords: teachers' assessments; summative; validity; portfolios; pedagogy

Introduction

This paper presents the findings of an exploratory and qualitative research that aimed to work with teachers of English and mathematics teaching pupils in the age range 12 to 16 in three secondary schools in England. The study aimed to address two questions:

- What would be the elements of a strategy which could enhance the quality of teachers' summative assessments in ways that would be both feasible and judged to be positively valuable by teachers?
- Would this strategy promote a positive interaction between formative and summative assessment practices?

The term 'quality' is used here in the sense in which some use the term 'dependability' (Mansell, James, and the Assessment Reform Group 2009) to comprise both validity and reliability.

In the UK, recent changes in the national systems for summative testing and accountability have given teachers more freedom, and more responsibility, in the summative assessment of their pupils. In Northern Ireland and Wales mandatory external tests at ages 7, 11 and 14 have been discontinued, Scotland has not

*Corresponding author. Email: paul.black@kcl.ac.uk

ISSN 0969-594X print/ISSN 1465-329X online

http://dx.doi.org/10.1080/0969594X.2011.557020
http://www.tandfonline.com

instituted mandatory external tests below the school-leaving age, whilst in England tests at age 7 have been replaced by a system based on teachers' assessments; national tests at age 14 were discontinued after 2008, but this was too late to affect the present study. These developments raise questions about the quality of teachers' assessments, and about the types of programmes for continuous professional development (CPD) which might enhance that quality.

A key feature in all such changes has been debate about the respective roles, for summative assessment purposes, of external tests and of assessments conducted by teachers. In England, the abandonment of most of the so-called 'coursework' assessment components of many of the GCSE[1] examinations in England (Whetton 2009; QCA 2006), the introduction in their place of 'controlled conditions' assessments (QCA 2006, 2010), and government promotion of resources and regimes for frequent summative testing (QCA 2009; DCSF 2008; Stanley et al. 2009) are all signs of an uncertain situation.

One aspect of this debate has been concern about the dependence on summative tests, in respect of their validity, their reliability (Stobart 2009; H of C 2008), and their harmful effects on teaching and learning (Mansell, James, and the Assessment Reform Group 2009). To these have been added arguments that a common assumption that teachers' assessments cannot be trustworthy is unjustified, quoting evidence from research studies to spell out the conditions required for such assessments to achieve their potential quality (Harlen 2004; ARG 2006; Daugherty 2010). Several national systems use teacher assessments as part of their high-stakes test, but, as Black and Wiliam (2007) pointed out:

> . . . while many systems rely on teacher judgment for assessments that are high stakes for students, there are . . . no systems that rely on teacher judgment for assessments that are high stakes for teachers. (11)

Gardner (2007) argued that one indicator of the low status of the profession was that, unlike (say) medicine or the law, it does not have control over the assessment dimension of its task. He argued that 'a general milieu of distrust' is one reason, and that another is that the literacy, the skills and the values of the profession in relation to assessment are all weak. In what follows, we shall use the term 'assessment competence' to represent the combination of literacy, skills and values which is set out in Gardner's article.

This discussion shows that the present work's focus, on improving the quality of the summative judgments made by teachers and schools, may be seen as serving two purposes. The first is that these judgments, in the context of each pupil's progress through the school, affect advice about options exchanged between teachers, and with the parents and the pupil. The second is that such judgments can be used to contribute to national assessments of pupils and schools, a contribution which can only be justified if those judgments are well founded and based on procedures and criteria which are comparable within and between schools. Whilst procedures for the latter purpose clearly require the processes variously denoted as calibration, standardisation or moderation, we assumed from the outset that some such procedures would also be required for the first purpose because the commitment to aligning procedures and criteria would give essential impetus to the necessary process of professional development through collegiate collaboration. Thus the development and use of moderation exercises was a key part of the work described below. Whilst

teachers were familiar with moderation requirements developed at a national level, the fact that teachers had in general supported the abandonment of coursework assessment in mathematics at GCSE (QCA 2005) showed that these were not fit for their purpose. Practices in English at GCSE were more popular with teachers and influenced our work. Overall, the 'coursework assessment' situation was fluid because of the forthcoming 'controlled conditions' innovation (QCDA 2010). However, in neither case were these national requirements designed or used for the year-on-year summative assessments in secondary schools, and there is controversy about whether the coursework component of the assessments at age 14 improved the quality of teaching and learning.

In the next section we describe in turn the procedures, evidence collected, and analyses used in the research. Two subsequent sections give an account of the main findings, and a final section will review these, comparing them with what has been established in the existing literature and considering their implications for future developments in policy and practice.

The King's Oxfordshire Summative Assessment Project (KOSAP)

This project was a collaboration between ourselves and the Oxfordshire Local Authority (see Black et al. 2010 for full details). It involved collaboration with 18 teachers over a period of two-and-a-half years, drawn from both the English and mathematics departments of three schools.[2] An audit of their existing summative assessment practices led to interventions to encourage changes aimed at improving them. All three schools worked on the assessments of students in year 8 (ages 12–13).

Between March 2005 and November 2007, all participants took part in nine whole day meetings. The research data included classroom observations, records of all relevant meetings, transcriptions of individual and group discussions with teachers, reflective writing by the teachers, and the assessment tasks developed and used in the course of the project.

In this paper we draw mainly on the teachers' views, as expressed in the interview transcripts and in their reflective writing. These texts were analysed, using a coding scheme partially derived from theory and partially grounded in the data (Glaser and Strauss 1967). Reliability in the application of agreed codes was cross-checked between pairs of team members. Those sections in the texts which were concerned with assessment competence were selected (250 paragraphs in all) and grouped according to the different issues raised.

In what follows, the term 'tests' will be used to represent procedures in which pupils attempt a set of questions, previously unseen, with limited time and working entirely on their own. Such a test may be external, one in which the school has no say in its formulation and marking, or internal, if the school selects and assembles the questions and marks the responses. The term 'task' will refer to assessment activities in which pupils may have notice of and preparation for the production of the work to be assessed, collaboration may be encouraged, and the conditions may be more similar to those of normal classroom work.

The initial audit

Our approach was to undertake an initial audit, and then in the light of that to identify the work that would be needed to achieve our aims. A first finding was that summative practices varied between different teachers and different subject

departments in the same school, and between schools. A detailed account of the audit results is given in our previous paper (Black et al. 2010): the account here selects and highlights some features which were particularly relevant to the aims discussed in this paper.

Teachers thought that school managements were mainly interested in the results of external tests, and of similar tests which, in intervening years, could help to predict the results of the external tests. Thus whilst schools were free to devise their own summative assessment practices in most of the five years of compulsory secondary education, the practices were strongly influenced throughout by external tests.

Indeed, because of the focus of both parents and school managements on external test scores, the teachers' reputations depended mainly on these scores, so that they felt that they had to prioritise preparation for the tests, even where they saw this as lowering the quality of their teaching. There was little to motivate teachers to develop their own assessments. In consequence, their pupils experienced teaching influenced by the negative effects of the discontinuity between formative and summative practices, and suffered from the stress that tests created for them (see Harlen and Deakin Crick 2002; ARG 2002).

One outcome was that teachers did not develop their skills at composing test items, since their best strategy was to copy the external testing regime. The abandonment of the coursework component in mathematics at GCSE seemed to have confirmed the habit, in mathematics, of reliance on the use of items from the external tests in all years. However, in English teachers had a stronger tradition in conducting non-formal assessments and were committed to the system for GCSE coursework. This work was nevertheless hindered by the specifications of the national curriculum levels:

> They're just not helpful. … [I] mean they are kind a vague verbalisation of gut feeling. They are inconsistent within themselves. They don't cover the full range of skills that we are attempting to assess. (English teacher)[3]

In this situation, one main starting point of the project was to engage the teachers in fresh consideration of the concept of validity. This aspect of the work has been described in detail elsewhere (Black et al. 2010). It was central to establishing the orientation of a renewed summative assessment system, as will become clear below. It is only necessary to point out here that teachers recognised the weak validity of their existing practices. Outstanding in the case of English was the absence of assessment of speaking and listening, whilst one mathematics teacher described the situation as follows:

> And we found that we were testing them to get them to be good at doing tests really. And we wanted there to be more to it than that. (Mathematics teacher)

The main findings: five steps to improvement

Our main findings are best described in terms of the five steps described below. This structure, and many of the detailed issues within each step, were not planned beforehand, but emerged as we learnt, from the initial audit and then in the collaborative development with the teachers, both to refine our view of quality and to clarify what was needed to achieve it.

The design of tasks

The initial focus of our work was to help the teachers to design and then improve tasks to be implemented and assessed by them in a variety of classroom conditions. A second focus followed later when it was envisaged that each pupil's summative assessment might be based on a portfolio of task assessments together with test results.

Many teachers already used realistic tasks which pupils could tackle in normal classroom conditions and which gave opportunity to develop creativity and flair in the use of language and in the application of the skills they were learning. We were able to add to their store of potentially suitable tasks, but it was up to the teachers, working together, to select and refine these. The refinement was iterative, in that some shortcomings in a task only became evident when it was used in practice. One of the key lessons learnt was disclosure, i.e. that a task should evoke evidence of the attainment it was designed to test:

> . . . an essay tends to be all about their writing and weaker students just rely completely on a writing frame . . . the writing frame tends to be structured, structured to show their skills of inference where in fact you can't even see where they've got skills of basic retrieval in that essay. That's often done for them in the writing frame. (English teacher)

In the work of improving tasks, our initial focus on engaging teachers in exploring validity paid dividends. A further key lesson emerged when pupils' reports of their work showed that in their initial form, some tasks were beyond some pupils so that they produced nothing of merit, whilst other tasks had too low a ceiling so that the strongest were not stretched.

A corollary here was that a task should engage and motivate the pupils. A mathematics task about the relationship between the standard sizes of paper had this quality:

> I think one of the reasons is that it's a good task is that it's a real task. It's all based on A4, A5 and A6 you know, which is in real life, kids know that. . . . they can physically hold up an A5 sheet against an A4, . . . a lot of kids are engaged straight away. . . . it's a piece of work that every kid could achieve from. (Mathematics teacher)

It is hard to go further here without discussing a variety of specific examples. Overall however the teachers agreed that the opportunities provided to work together on task construction had both improved their judgment of task quality and produced practically useful teaching materials.

Implementation in the classroom

Pupils' performance on a task will depend in part on the way it is implemented. Three issues arose under this heading, as follows:

- The clarity of presentation of a task was a necessary condition, but not sufficient because if pupils had no experience of tackling similar tasks, many would not know what they were supposed to do:

> . . . at the end of units, they'd all be really nice investigations to do that would allow them to think about things that they've covered in class, but also, you know, be

> exposed to the investigations of how to do them, what to think about. Because I was quite surprised at the beginning how they couldn't ... they didn't know how to do it. (Mathematics teacher)

Closely supported work on preliminary tasks, conducted exclusively for formative purposes, and exploiting the advantages of using peer assessment for developing pupils' understanding of the relevant aims and criteria, was often needed.

- Each teacher had to decide to what extent pupils should work together, how much help to give different pupils, and how to record such interventions. Teachers felt they should give help to save some pupils from 'hours worth of meaningless exploration'. In the case of an exercise designed for inclusion in a summative assessment portfolio, this was justified where teachers could make allowance for such help. However, they saw the need to draw a fine line between giving help, and merely responding to pressure for help from pupils who wanted to be told what to do and to evade the need to think things out for themselves.
- Because moderation of assessment standards within and between schools was envisaged, there had to be guidelines to secure uniformity of practice across all the schools; some teachers were concerned that without such guidelines, a few would 'cheat' by giving excessive help. However, even with honest attempts at fair presentation, problems could arise:

> ... I remember [colleague] and I doing the same task and obviously introducing it very differently and getting very different results. So I think having some agreed starting point is essential. (Mathematics teacher)

Another threat was variation in the amount of help pupils might be given outside the classroom. In cases where plagiarism or copying might be involved, teachers judged that they knew their pupils well enough to be able to detect such malpractice. More subtle problems arise where extra support has actually led to improved learning, for then it would be fair to assess the outcome although some had enjoyed more support than others in achieving it. To offset this, teachers would help by ensuring that all had access to the essential resources, by including some work produced under controlled conditions, and by using their knowledge of the profile of performance of any pupil to detect and explore discrepancies (see also Stanley et al. 2009, 71).

Portfolios – balance, range, content

Any one task could not validly reflect all of the aims in either subject, whilst in addition the fact that pupils might be assessed on several occasions in a variety of contexts should enhance dependability in several other ways. Thus, a collection of performances on several tasks in a portfolio was assumed to be necessary, but this general rationale had to be followed up by a suitable choice of the components.

Within-school and between-school meetings discussed the composition and selection of the portfolio components: the validity of summative assessments was seen to depend on the range and balance of the contents of each pupil's portfolio, in that these contents should reflect the range of aims of each subject and should be varied in style. The inclusion, in English, of separate pieces of evidence about writing, reading, speaking was an example of improved range:

> So, I feel like focus in the past has been largely on their writing. And I think the result of this project would be they would comment more on their reading skills and on their speaking and listening. I think it will mean a much you know, a much richer and wider report. (English teacher)

However, more changes were still to be explored. Examples were:

- the assessment of speaking and listening was still seen as problematic;
- several aspects of the subject, such as poetry, had not yet been considered;
- more varied tasks to broaden the range of opportunities for the pupils;
- achieving flexibility in matching tasks to the interests and ability of pupils, but with all assessing the same aim.

There was a similar concern for auditing the range of tasks used in mathematics: as one teacher put it, to 'cover all aspects of a varied and balanced diet'. Some wanted to explore the use of oral assessments, particularly with pupils of whose assessments they were not confident. One school was looking at the contribution of formal tests; at the end of work on a topic, there might be an open-ended classroom task and a formal test, both tailor-made to give as full a picture as possible of the learning achieved.

Overall it was clear that more development was needed to adjust the components of portfolios so that they were linked to the teaching and gave a comprehensive and balanced picture of pupils' achievements.

Marking and aggregation

The next step in strengthening the dependability of the summative results lay in securing a shared consistency in the interpretation of assessment criteria. Teachers felt that the criteria available from the national curriculum were not adequate guides for consistent judgments. This was particularly evident in the using and applying of mathematics. When assessing pupil investigations in particular, the mathematics teachers expressed a need for the criteria to be broken down into sub-level criteria as was the case in national testing. The levels for Using and Applying in Mathematics are somewhat vague, particularly in the use of qualifiers such as 'simple' and 'quite complex'. However, it appeared to us that the problem was more related to the teachers' inexperience in direct use of the National Curriculum. This was partly due to institutional expectations: schools' policies called for regular tests which had to report results in terms of sub-levels, which could be achieved by using ready-made test items and their mark schemes and reporting on the proportions of questions answered correctly. Assessment was thereby separated from curricular understanding and tests became a collection of questions from past examination papers or textbooks. Such a system, whilst easy to administer, did not fit with the teachers' desire to foster a formative approach that guided pupil involvement in their learning.

The teachers were rather more comfortable implementing the criteria for GCSE coursework, with which they were more familiar. On the other hand, these detailed criteria were found to be restrictive and to reward formulaic approaches to investigations. Indeed, the teachers recognised that very tight descriptors might discourage pupils' creativity and promote routine over non-routine approaches.

Task reports were found more difficult to assess than test papers. In mathematics, some pupils produced ample material, but careful scrutiny was needed to

distinguish reports which showed understanding from those which mentioned the appropriate terms and techniques but did not relate these to the task in valid ways.

Similarly, creativity and flair in a writing assignment could not be judged by an analytic approach – the project helped teachers to rely more on their holistic judgments:

> Project has removed anxiety about delineating success only in terms of a neat, prescriptive check list. (English teacher)

However, such judgments had to be defensible by reference only to the text. Whilst a teacher who knew a pupil well could read more into that pupil's work than any external assessor, a moderation group could only base their agreed judgment on the text alone. It followed that a portfolio should be so composed that the collection gave a self-contained picture of the pupil's achievements.

It was found hard to formulate agreed ways to arrive at an aggregate assessment of a portfolio's collections of diverse pieces of evidence. One aspect of this problem was the variability in the outcomes for some pupils, both between different classroom tasks, and between these and the results of formal tests. It became clear that different teachers tackled such problems in different ways: some suggested that, whilst flexibility might be needed, a form of flow chart might be a helpful guide to such decisions.

Standardisation and moderation

As explained in the Introduction, involvement in exercises for the alignment of procedures and criteria, within and between schools, was an essential component of our strategy for improving schools' own summative assessments, and these would also provide occasions for teachers to learn through discussion with one another. The strategy here could involve either a Standardisation or a Moderation approach. Standardisation involves training on a common set of examples, after which schools are assumed to be competent in applying common criteria and standards to their own work. Moderation is more rigorous in requiring both intra- and inter-school meetings every year with blind marking[4] of real current examples. The English teachers, being familiar with a standardisation approach for GCSE, agreed on the value of using standardisation materials, but some thought that, after this, blind marking should not be needed in the exchange of current samples before the project's moderation meetings. The mathematics teachers, who had not experienced this approach, were content, at the stage reached in the project, to operate with the full moderation procedure.

It was a surprise to some teachers to discover the wide differences between the judgments of colleagues about the same samples of work, as exposed at the moderation meetings. In some departments, comparability exercises of this type had not previously been deemed necessary. One cause of difficulty has already been mentioned – teachers' perceptions of the inadequacy of the available criteria. It was predictable that, where the gaps created by vagueness had to be filled by personal judgment, rather than by shared and negotiated interpretations, differences would ensue.

The meetings had also exposed the problem that some teachers interpreted the discussion of differences as an attack on their competence:

> I think that raises issues, you know personal issues about people feeling that their judgement is being questioned or being undermined or feeling that there are people in that public setting who are questioning what they are doing. (Mathematics teacher)

On the whole, these exercises were seen to be very valuable: they had helped individuals to re-think their own understandings of criteria and standards, and to be more confident about the validity and consistency of their judgments, in ways that would have a positive impact on the achievements of their students:

> . . . that the moderation and standardisation process was incredibly valuable in ensuring rigour, consistency and confidence with our approach to assessment; that teachers in school were highly motivated by being involved in the process that would impact on the achievement of students in their classes (like the moderation and standardisation at GCSE). (English teacher)

> And we've had moderation meetings, we were together with the other schools, teachers in other schools looked at how rigorous our assessment would be and they criticised what, you know, our marking criteria is. And we changed it, which has all been very positive. (Mathematics teacher)

In one school it was planned to have such meetings three times a year. Some were considering the possible value of meetings to 'moderate formatively', i.e. to explore ongoing progress in developing shared judgments and criteria in advance of the occasions where decisions would have to be taken. Because the broader focus of their assessment work gave more opportunity for pupils to perform, the results revealed more to their teachers about ongoing possibilities of improving engagement and motivation:

> . . . you could see quite a lot from what people do, from how much work they put in outside of the classroom. . . . and you can see quite a lot about how they think, as well. And how they work in groups is quite interesting . . . (Mathematics teacher)

The impact of the project

Initially, the main impact of the project was in the development of assessment competence, firmly related to validity, by those teachers directly involved. The findings about this aspect are presented in detail in the previous section. This section presents the reports, by these teachers, of their views about the wider impact on three groups, namely their pupils, the pupils' parents, and their colleagues in each school. Direct evidence on these matters was not collected from any of these three groups.

Impact on pupils

Teachers perceived that many pupils saw teachers' assessments as a 'guess' – to be confirmed by the test results. They had experience of teachers whose approach to assessment was within the 'tick-box' culture:

> I mean the kids, the kids would have to be educated in this is a different way of doing things . . . They come to us with expectations, those same expectations that do exercises and teacher marks the exercises, tick, tick, tick, tick. Because that's what happens at primary school. (Mathematics teacher)

Teachers' awareness of the need to engage pupils so that they understood the aims of open-ended tasks led to new emphasis on explanation; for example, when a

mathematics class were asked to formulate and explore a hypothesis it became clear that many did not know what 'hypothesis' meant. More generally, as it became clear to pupils that this work differed in character from normal classroom work:

> I think it changed the dynamic of the lesson a little bit, in terms of well, in terms of there being much more an element of them getting on trying to find out . . . they were trying to be more independent, I think, I think some of them struggled with that, and others . . . some of them, some still find it quite difficult if they are not hand held all the way through when others were happier to sort of, go their own way. (Mathematics teacher)

The enhanced commitment so produced could become problematic. For some pupils who had invested a great deal of effort in their first tasks, the enthusiasm began to fade when the next task was presented, even though the initial degree of effort had been neither required nor expected.

The use of peer-assessment discussions in which pupils were asked to rank a set of samples of work from previous years seemed to help to develop their understanding of the aims and criteria, as one English teacher put it, 'so they know what to aspire to', whilst a mathematics teacher commented that:

> I think the strength is that it's genuine, it's much more, it's much better for mathematics, I think much better for life. How have you thought about this, what's your solution to this, how else could you have done it, what other angles would you consider, what were the multiple answers you got? Rather than you got the right answer. . . . this is much more satisfying. (Mathematics teacher)

The work also encouraged the lower attaining pupils, and helped all to begin to see the progress in their own achievements:

> . . . the kids felt it was the best piece of work they'd done. And a lot of teachers said 'OK we're really pleased with how they've come out', it's seemed to really give kids the opportunity to do the best that they could have done. . . . Err, it just, the lower end kids it organises their work for them, it's basically a structured path for them to follow. (English teacher)

The overall benefit experienced by the teachers was that pupils began to see this summative assessment work as a shared enterprise. For evidence to take home, the 'working-at-grade' report was replaced by their portfolios which they regarded as theirs, and which they could describe to their parents.

Impact on parents

Previously, teachers had used test results to tell parents what most of them wanted to know: their child's level and a prediction of future test results. With portfolios, teachers felt more confident:

> But I think if all the teachers had more, possibly more ownership of what we are actually doing in terms of summative assessment then you would have more confidence in saying to parents, which I think is one of the biggest things I find with lower school. (Mathematics teacher)

The portfolio evidence gave greater credibility to the teacher's assessment, partly because it provided, for discussion, a richer view of a pupil's achievements:

> It provides you with convincing, comparative material and it's something that I have taken to meetings, with parents of students in classes other than my own. Being able to say, 'well this is what John's been able to do and this is why he is still on a 5C and not a 6A'. And parents seem to react well to that ... (English teacher)

Some felt that this function of communication could well be explored further, for example by adding a summary in terms of a profile across the several attainment targets linked to the portfolio's contents.

Impact on colleagues/school

A first stage in the teachers' process of change was the need to convince sceptical colleagues, by challenging pre-conceptions about assessment and showing that the use of portfolios was feasible in practice:

> Implementation of tasks – staff were a bit reluctant to do the projects ... but post projects their views changed and this year development of investigation-based tasks has become an issue that the KS3[5] staff have been keen to do and is being done as part of performance management. (Mathematics teacher)

The conversations that were provoked were seen as fundamental in raising issues about achievement criteria, about the quality of pupils' work, and so about important aspects of teaching and learning:

> I think it's quite a healthy thing for a department to be doing because I think it will encourage people to have conversations and it's about teaching and learning. ... it really provides a discussion hopefully as well to talk about quality and you know what you think of was a success in English. Still really fundamental conversations. (English teacher)

The collection of examples, of tasks, of work of pupils, and of their assessments, would help when new staff had to learn about the summative assessment policy.

The portfolios could provide teachers with a wealth of information to communicate to those who would teach the same classes in the subsequent year. However, some doubted that teachers would have time to absorb such information, arguing that if pupils' performance on the tasks were part of their pupil record there was no need to look at portfolios.

However, the project's renewal of emphasis on summative assessments was seen to contribute strongly to communication between colleagues; as one teacher put it:

> ... everybody in this department seems to do it in a very different fashion at the moment and it would be nicer to do it in a way that meant that we are all singing from the same hymn sheet if you like. I think it's made us think very carefully about how we construct our year and construct our year span in terms of what comes when and what's being assessed when. (English teacher)

Achieving our aims

Answering our first question

Our first question was:

- What would be the elements of a strategy which could enhance the quality of teachers' summative assessments in ways that would be both feasible and judged to be positively valuable by teachers?

Our initial view, that we should first help teachers to debate validity and then contrast present practices with their own conclusions from this debate, was amply confirmed. It helped to establish that change was needed. They saw that they had to look at new ways of learning about their pupils' capabilities, and that they had to focus more on teaching certain skills and on assessing some important aspects of their subjects that had been overlooked. The five steps, as described above, all followed. However, the work has highlighted several critical features of these steps.

The most important conclusions relate to the nature of what is assessed in each subject. This research suggests strongly that in English assessment had to be designed to give separate evidence of writing, of reading, and of speaking and listening, whilst in mathematics they should assess the capability to apply mathematics skills to authentic and non-routine problems. Our research highlighted ways in which existing assessments discouraged these aspects of the curriculum and, in doing so, seriously de-skilled teachers. A related and equally serious obstacle to the matching of assessments to aims was the ways in which teachers understood existing National Curriculum criteria. However, we emphasise that this is not simply a matter of extending the content or contexts of assessment. Rather, it relates to the nature of school English and mathematics and the nature of teachers' relationships, experiences and knowledge of the disciplines. The teachers only became aware of this problem when they took on responsibility for designing tasks which met their own criteria for validity and then for marking these in ways that would command agreement between colleagues.

The need to strike a balance between uniformity and diversity was another outstanding problem. It seemed that their agreed system had to specify five main features in respect of which there had to be some degree of uniformity if moderation, both within and between schools, were to be feasible:

- The validity of each component of a portfolio. Each task or test has to be justified in relation to the aims that it was said to assess.
- Agreed conditions for the presentation and guidance under which pupils would work in producing the various components of a portfolio. This problem was exposed and discussed, and some ground rules, to balance uniformity with flexibility, were agreed. There was hope that the combination of agreed controlled conditions with strong moderation systems could alleviate the potential problems.
- Guidelines about the ways in which each domain is sampled within a portfolio. Some holistic portfolio tasks can cover a wider range of knowledge and skills than individual test questions, but it also emerged that some attainment targets could only be assessed by tasks of limited range. The teachers did discuss this problem, but it would take more cycles of assessment to inform the optimisation of this aspect. It was not envisaged that a small number of open-ended tasks could entirely replace multi-question tests, rather that both types should be included in any portfolio. However, the work did not advance far enough to develop this aspect. One feature of the agreed procedure was a compromise between uniformity, requiring some tasks to be common across schools, and a degree of flexibility in the choice of the others.
- Clear specification of the criteria to which all assessors have to work. Teachers need time to develop shared understandings of any criteria, and institutional reporting requirements need to be aligned with these criteria. Particular

difficulties may arise if teachers have to work to more than one scheme. Thus, the removal of age 14 tests, so that only the GCSE criteria were important, should be a helpful development. Standardisation exemplars, with guidelines for assessment procedures, for use in training in inter-school moderation, would be needed.

- A requirement for comparability of results within and between schools. The attention given to this aspect helped motivate collegial discussions between teachers, which enriched the feedback needed to develop assessment skills. However, whilst the moderation procedures would have helped to improve the reliability of the assessments, we did not attempt any systematic check on this aspect. The need to resolve the differences in judgment encountered in the moderation meetings was an important part of the learning process of the project, but the outcomes of such meetings could not provide evidence of the possible agreement, within and across moderation groups, in a developed system.

All of the above conditions bear upon the overall validity and reliability of the assessments produced, forming a linked chain similar to that set out by Crooks (1996).[6] Whilst validity was a key criterion in the development of the components and of the overall composition of the portfolios, the task of exploring the limitations of, and the threats to, the reliability of the outcomes, which are implicit in the above conditions, could not be explored in this project. The approach explored could be robust in sampling over several task occasions, distributed over different times, and attempted in conditions quite different from those of formal tests.

That the project's outcomes were seen as feasible and rewarding by the teachers has been amply attested in their interviews and in their reflective writing – although they were not claiming that most of the problems had been resolved. The teachers all felt that extensive professional training would be needed if others were to learn both the principles and the procedures that were central to this project:

> I think the department will need to go through the sort of thing that we've gone through, but obviously a little bit speedily or speeded up. So that thinking about what makes a good mathematician; the thinking about the tasks before you give them to the group; and thinking about the criteria, because I think all those are valuable routes to eventually being able to moderate the task. (Mathematics teacher)

> But I think it would be essential if everybody had clear training and … how the portfolio would look, what the tasks … would look like. Obviously samples, portfolios you would want, wouldn't you. You would get a sense of what, what task would be appropriate … I think you'd have to that, otherwise you are going to get teachers going 'I don't know what I'm supposed to do.' (English teacher)

It is clear that a programme of this type could not be 'rolled out' through use of printed materials backed up by a brief CPD session. Rather, what would be needed, as stated above, would be a 'speeded up' version of this project, implemented through organisation of local groups of schools.

The above findings do identify problems that will need attention if the quality of teachers' summative assessments is to be improved. The work established, in the opinions of the teachers, and of ourselves as subject experts, that their validity had been enhanced. It could not, within the time limits of the developmental work,

produce any measures of the reliability of possible outcomes of the procedures developed. Whilst the work was planned with such measures in mind, it turned out that the extent of the collaborative development, between ourselves and the teachers, that was needed was such that we could not get beyond the development, with the teachers and schools, of the skills and practices needed to achieve this aim.

Answering our second question

- Would this strategy promote a positive interaction between formative and summative assessment practices?

In work described by Black et al. (2003), one aspect of this problem was tackled by promoting the formative use of summative tests. The approach in this project complements this, going further in embedding the tasks for summative assessment within formative classroom work. The key findings here were set out in the sections on *Use in the classroom* and on the *Impact on pupils.* The teachers were better placed to achieve a positive interaction because they were in control of both the formative and the summative procedures, including control over the shaping and implementation of the assessments together with confident 'insider' knowledge of the procedures for the appraisal of pupils' work. All of this is in sharp contrast to their lack of power and understanding with external tests, which leaves them without the freedom of manoeuvre to forge a constructive interaction.

A helpful feature was the two-stage process, with a focus on pupils experiencing and learning about what a task required in the first stage, where formative interaction had priority. In a second stage, there would be little or no support and it was the product of this stage that went into the portfolio. The tasks in the two stages could be the same task, or two closely parallel tasks. The aim was that all pupils would have a clear idea about what was expected of them. The section on *Impact on pupils* highlights this advantage, for it shows that the informal and staged nature of the process makes it possible for pupils to understand and become more involved in the summative assessments, whilst also being more actively engaged so that they did themselves justice.

This staged process is similar in some respects to the new form of 'controlled conditions' assessment recently established as a replacement for 'coursework assessment' in England's GCSE examinations in some subjects, not to date including English and mathematics (QCA 2010; AQA 2010). Proponents of dynamic assessment (Poehner and Lantoff 2005) argue that a pupil's production on a task, after having been given a great deal of feedback and other help with a first attempt, is a more valid product for assessment than one that is done 'cold'.

Implication of the findings

Given that this was an exploratory study with a small sample, any implications should be treated as guidance for further work rather than as definitive. However, we point out that although the schools were chosen because of their success in the Oxfordshire Authority's work on developing formative assessment practices, the findings of the initial audit indicated needs, for the development work we undertook, which were more extensive than we had envisaged. We infer that the many problems that have been discussed above would need attention in any programme

aimed at improving assessment in schools, even though we cannot claim that we used the optimum approaches to tackling them.

Broader implications for schools

The summative assessments of their pupils which schools make year on year have an importance, for all of the stakeholders, independent of the various impacts of externally imposed assessment systems. So the quality of these assessments is important and this study goes some way to mapping out the work needed to audit current practices and work for the improvement of teachers' assessment competence. The sequence of an audit, followed by development in the five steps set out as our main findings, taken together with the discussion so far in this section, could serve as the agenda for such work. In particular, it sets out ways in which tension between formative and summative requirements can be tackled, at least in those school years for which schools have control over their own summative practices – providing they have the courage to exercise this control and take responsibility for their own ways of enhancing pupils' achievements. There remains however the larger question of the implications for any national assessment system.

Policy implications for national systems

The findings above raise, and could contribute to, debate about whether a dependable system could be established on the basis of a portfolio approach. Studies which set out to explore the reliability of teachers' assessments of student portfolios have produced diverse outcomes but serve both to indicate some of the conditions for achieving dependable results and to highlight some of the threats to reliability which are implied in the main features set out in the discussion above of our first main aim. A common theme is the tension between flexibility and comparability, as identified by Koretz (1998):

> Portfolio assessment has attributes that make it particularly appealing to those who wish to use assessment to encourage richer instruction – for example, the 'authentic' nature of some tasks, the reliance on large tasks, the lack of standardization, and the close integration of assessment with instruction. But some of these attributes may undermine the ability of the assessments to provide performance data of comparable meaning across large numbers of schools. (332)

Whether the portfolio approach is the best way to resolve this tension is an open question: Shapley and Bush (1999) give a more hopeful account of a portfolio development in a US state. Any system which was more loosely defined, such as one relying on informal collection of evidence noted during normal classroom teaching, would inevitably meet the obstacles outlined above and would find it more difficult to deal with them. Thus it would seem that the use of a collection of products of performance on a set of tasks, more or less closely specified, would be an essential feature of any dependable system. To this must be added the advantages of using such collections in inter-school moderation exercises to improve and align school practice. The value of such exercises was argued, and researched in practice, in the Scottish context by Hayward, Dow, and Boyd (2008).

However, it is not possible to draw general conclusions from the results of other studies of teachers' summative assessments because each such result must depend

upon the details of the system being studied. For example, studies which show that teachers can be trained to be reliable markers of the same pieces of work (Good 1988), or which explore reliability if such training is combined with a constraint that all use the same externally prescribed task according to conditions of uniform presentation (Taylor 2005), may be hopeful indicators: however, they do not allow for some of the flexibility which was found to be an important feature of the present project. If a system were based on specifying tasks and tests externally, it might lead to a regime of frequent testing, or of stereotyped tasks, which might lose those features of validity which a system should be set up to secure. A system combining external tests with components assessed by schools has obvious attractions, but then it would be necessary to consider why this type of solution, as operated in England's GCSE for over 15 years, did not command respect for its validity, and was abandoned for many, but not all, school subjects.

A more hopeful example is the systems developed in the Australian states of New South Wales and Queensland. A detailed account of these is given by Stanley et al. (2009) in their review of several national systems. What their account makes clear is that the systems developed, and well established, in these Australian states specify procedures which assess a limited number of tasks from each pupil (i.e. portfolios), with the state providing extensive resources for professional development including resource materials of many kinds. Moderation procedures involve either scaling against external tests in New South Wales (where tests and teachers' assessments each contribute 50% of the total), or joint work in school groups in Queensland where outcomes are based entirely on teachers' assessments. What is significant here is that many of the details of the operation of these systems reflect the steps described in our study.

The work of our project has spelt out some of the conditions that must be met if a dependable system were to be developed, because it has made considerable headway in tackling many of the problems that other systems have encountered. In addition, it has been seen as acceptable, and valuable to the teachers involved. We would also claim that it has established guidelines which if applied now would improve the quality of the internal summative assessments of many schools.

It is clear however that further exploration is needed, to develop a fully fledged system and to produce evidence of the validity and reliability that could justify the use of such assessments in public examinations. Any such development should include analysis of the differences between the needs of different school subjects, and of the prospect that any new system would suffer from the corrosive effects of 'teaching to the test' with inevitable stereotyping in any of the assessment tasks involved. At a different level, there would have to be a strategy, agreed between the profession and government, to tackle three obstacles: the need for investment in new professional development and in new types of assessment resources, teachers' concern about enhanced workload, and the public understanding of the issues involved.

Acknowledgements

We are grateful to the subject teachers in the three schools, and to their senior managements, for the willing collaboration on which this project was based. We also acknowledge the help of Dorothy Kavanagh and her colleagues in the Oxfordshire Local Authority for helping to set up the project and for guidance and support throughout, both to the teachers and to ourselves. Finally, we acknowledge initial funding by the DfES and subsequent support from the Nuffield Foundation.

Notes

1. GCSE: the General Certificate of School Education – subject examinations taken in England, Wales and Northern Ireland by most pupils at age 16, the end of compulsory secondary education. The changes had little effect on the teachers involved here as they were due to happen after, or near the end of, the project.
2. There were at least two teachers from each school department, and there were some changes and substitutions made by the schools during the course of the project.
3. Quotations ascribed to teachers are given exactly as expressed in their oral contributions in interview, or as written by each.
4. Blind marking is a procedure in which each marker sees the work with no marks or comments added, and assigns a mark without consultation with others; these independent marks are then tabled at the outset of a meeting between the markers.
5. Key Stage 3: the first three years of secondary education, ages 11 to 14.
6. Crooks, Kane, and Cohen (1996) treat reliability as one of the components that determine validity. In this paper, we treat reliability and validity as separate components of dependability or quality.

Notes on contributors

Paul Black is Professor Emeritus at King's College London. His main contributions have been to curriculum development in science and in technology, and to assessment research. He chaired the government's 1998 Task Group on Assessment and Testing and also served on advisory groups of the USA National Research Council. He was a member of the Assessment Reform Group. His recent work with the King's Assessment for Learning Group has been concerned with formative and summative assessments by teachers.

After 13 years teaching in secondary schools, Christine Harrison joined King's College to run the Biology Education section. Her teaching and research have centred on assessment, science education, cognitive acceleration and the use of text and TV in classrooms. In 1998, she began work on the King's Medway Oxfordshire Formative Assessment project (KMOFAP), helping science and mathematics teachers to focus on and improve their formative practice. This led to other projects and consultancies in the area of classroom assessment. More recently, Chris has worked on teachers' professional learning, and on classroom research with a focus on the use of portfolios and diaries as reflective professional tools.

Jeremy Hodgen is Senior Lecturer in Mathematics Education at King's College London. His research interests include mathematics teaching and learning, assessment and teacher education. He is currently directing Improving Competence and Confidence in Algebra and Multiplicative Structures (ICCAMS), which is investigating how teachers can implement classroom assessment in mathematics.

Bethan Marshall is a Senior Lecturer in Education. She specialises in issues relating to the teaching of English and assessment. In addition to being part of the King's Oxfordshire Summative Assessment Project (KOSAP) team, she was, for two years, a director at King's of the Learning How to Learn project, funded by the Economic and Social Research Council. She has written extensively on the subject of English and assessment including her book *English Teachers: An Unofficial Guide* and as a co-author of *Assessment for Learning: Putting It into Practice*. She was the chair of the Liberal Democrat commission into primary education.

Natasha Serret was recently awarded a PhD for work that explored the cognitive potential of talk in primary science classrooms. Since joining King's College London as the senior researcher for the primary Cognitive Acceleration in Science Project, she has continued to work with teachers on approaches that emphasise cognitive acceleration, assessment and social interaction in learning.

References

AQA (Assessment and Qualifications Alliance). 2010. Assessment and Qualifications Alliance: Controlled assessment. http://web.aqa.org.uk/support/assessment.php?id=03&prev=03 (accessed February 3, 2010).

ARG (Assessment Reform Group). 2002. *Testing motivation and learning*. Assessment Reform Group. Cambridge: Cambridge University Faculty of Education. http://www.assessment-reform-group.org (accessed February 3, 2010).

ARG (Assessment Reform Group). 2006. The role of teachers in the assessment of learning. Assessment Reform Group. http://www.assessment-reform-group.org (accessed February 3, 2010).

Black, P., C. Harrison, C. Lee, B. Marshall, and D. Wiliam. 2003. *Assessment for learning: Putting it into practice*. Buckingham: Open University Press.

Black, P., C. Harrison, J. Hodgen, B. Marshall, and N. Serret. 2010. Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice* 17, no. 2: 217–34.

Black, P., and D. Wiliam. 2007. Large-scale assessment systems: Design principles drawn from international comparisons. *Measurement* 5, no. 1: 1–53.

Crooks, T.J., M.T. Kane, and A.S. Cohen. 1996. Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice* 3, no. 3: 265–85.

Daugherty, R. 2010. Summative assessment: The role of teachers. In *International encyclopedia of education*, Vol. 3, ed. P. Peterson, E. Baker, and B. McGaw, 384–91. Oxford: Elsevier.

DCSF (Department for Children, Schools and Families). 2008. National strategies: Assessment for learning. http://nationalstrategies.standards.dcsf.gov.uk/primary/assessment/assessingpupilsprogressapp (accessed January 12, 2011).

Gardner, J. 2007. Is teaching a 'partial' profession? *Making the Grade,* Summer: 18–21.

Glaser, B.G., and A.L. Strauss. 1967. *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.

Good, F.J. 1988. Differences in marks awarded as a result of moderation: Some findings from a teacher assessed oral examination in French. *Educational Review* 40, no. 3: 319–31.

Harlen, W. 2004. *A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes*. Research Evidence in Education Library. London: EPPI-Centre, Social Science Research Unit, Institute of Education. http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=61 (accessed March 20, 2010).

Harlen, W., and R. Deakin Crick. 2002. *A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review)*. Research Evidence in Education Library. Issue 1. London: EPPI-Centre, Social Science Research Unit, Institute of Education. http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=61 (accessed March 20, 2010).

Hayward, H., W. Dow, and B. Boyd. 2008. *Sharing the standard.* Edinburgh: Education Scotland. http://wayback.archive-it.org/1961/20100730134209; http://www.ltscotland.org.uk/resources/s/genericresource_tcm4579679.asp?strReferringChannel=assess (accessed July 2011).

H of C. 2008. House of Commons, Children, Schools and Families Committee. *Testing and Assessment. Third Report of Session 2007–2008*. *Volume I*. Norwich: The Stationery Office.

Koretz, D. 1998. Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education: Principles, Policy & Practice* 5, no. 3: 309–34.

Mansell, W., M. James, and The Assessment Reform Group. 2009. *Assessment in schools: Fit for purpose? A commentary on the Teaching and Learning Research Programme*. London: Economic and Social Research Council.

Poehner, M.E., and J.P. Lantolf. 2005. Dynamic assessment in the language classroom. *Language Teaching Research* 9, no. 3: 1–33.

QCA (Qualifications and Curriculum Authority). 2005. *New coursework arrangements for GCSE*. London: Qualifications and Curriculum Authority.

QCA (Qualifications and Curriculum Authority). 2006. National curriculum assessments for England. London: Qualifications and Curriculum Authority. http://webarchive.nationalarchives.gov.uk/20110223175304; http://www.qcda.gov.uk/qualifications/gcses/570.aspx (accessed July 2011).

QCA (Qualifications and Curriculum Authority). 2009. Improving skills for assessing pupil progess. http://www.qcda.gov.uk/assessment/82.aspx (accessed January 12, 2011).

QCDA (Qualifications and Curriculum Development Agency). 2010. Developing the curriculum, delivering assessments, reforming qualifications. http://www.qcda.gov.uk/qualifications/exams/3510.aspx (accessed January 12, 2011).

Shapley, K.S., and M.J. Bush. 1999. Developing a valid and reliable portfolio assessment in the primary grades: Building on practical experience. *Applied Measurement in Education* 12, no. 2: 111–32.

Stanley, G., R. MacCann, J. Gardner, L. Reynolds, and I. Wild. 2009. *Review of teacher assessment: Evidence of what works best and issues for development*. Oxford: Oxford University Centre for Educational Development; report commissioned by the QCA. http://www.education.ox.ac.uk/assessment/publications.php (accessed January 12, 2011).

Stobart, G. 2009. Determining validity in national curriculum assessments. *Educational Research* 51, no. 2: 181–212.

Taylor, M. 2005. *Teacher moderation systems*. London: Qualifications and Curriculum Authority.

Whetton, C. 2009. National curriculum assessment in England: How well has it worked? Perspectives from the UK, Europe and beyond. *Educational Research* 51, no. 2: 131–5.