Riding the interface: an exploration of the issues that beset teachers as they strive for assessment systems

Paul Black, Chris Harrison, Jeremy Hodgen, Bethan Marshall & Natasha Serret

King’s College London

In this paper we examine the issues and difficulties that arise when teachers attempt to develop and implement a system of summative assessment by teachers rather than by externally constructed tests.

There has been much recent controversy over coursework-based examinations with concerns centring on the issue of plagiarism. In mathematics, teachers’ concerns have extended to the validity of coursework. A recent survey of mathematics teachers overwhelmingly favoured the immediate abandonment of coursework at GSCE with the result that QCA have abolished GCSE coursework from this year onward (QCA, 2006). The decision for English has been put on hold, for now, but in the past English teachers have argued, often vehemently, for its inclusion in the way in which pupils are assessed. Coursework is still used for 40% of the GSCE and complaints occur every year about KS3 exams, which contain no course based assessment at all. On the other hand, there is evidence that highlights weaknesses in both the reliability and validity of current examination based forms of assessment (e.g., Black & Wiliam, 2006; Stobart, 2005).

It was in this context that we conducted our research into summative assessment by English and mathematics teachers at Year 8 (Y8). The King’s Oxfordshire Summative Assessment (KOSA) Project was designed to investigate ways of improving the quality of teachers’ summative judgements, to compare methods appropriate to the different disciplines, to explore the effects of summative assessment by teachers on pupils’ motivation and learning, to inform strategies for professional development in assessment practices and to provide guidance for trials and development on a wider scale. In an earlier paper (Black et al., 2006), we reported on our initial findings relating to the ways in which the teachers developed and engaged with summative assessment. And, while tentative, these helped both us and the teachers make sense of the trial phase and inform and support the later implementation. As in our last paper, we focus on the teachers: the quality of their judgement, the similarities and differences in the two subjects and their professional development. We are still conducting our data analysis and hence our findings are still tentative. However, we will begin to highlight some of the wider findings and implications that are emerging.

---

1 In the survey, conducted by the QCA, 68% of mathematics teachers favoured the immediate abandonment of coursework with a further 20% felt the data-handling coursework should be removed.
2 In an earlier consultation, 95% of English teachers felt coursework to be a valuable part of GCSE assessment (QCA, 2005).
**Summative Assessment By Teachers**

There have been several reviews of the efficacy of summative assessment by teachers, the most comprehensive recent study being the EPPI review (Harlen, 2004). A number of detailed findings emerged. Some findings, however, were difficult to interpret. They found, for example that where boys performed better in examinations in science and mathematics than girls, this may be because they were good at test taking or that they did not bother to invest in course based assessment. The review concluded that if course based assessment were to be introduced substantial training amongst teachers had to take place for it to be reliable.

Assessment practices in Australia show a slightly different perspective. Cumming and Maxwell (2004) have demonstrated that teachers, when well versed in the procedures of course based work, do show reliability in their judgements even when the stakes are high. The model for assessment in Queensland, which is wholly school based, aims for a criteria led assessment with ‘respect for teacher professionalism in judging student achievement’ (Cummings and Maxwell, 2004, p.93). This echoes the mode of assessment, finally abandoned in 1994 by the then Tory government, for the 100% course work examinations for English, by what is now the AQA. In mathematics, similar assessment systems largely based on coursework suffered the same fate.

Each of these systems relied heavily on active teacher involvement at all levels, thus building a community of teacher assessors. Both had first schools and then expert teachers moderating the pupils’ work. The AQA, which around 80% of schools in England take, have maintained their system of standardising and moderation to the present day at GSCE, although the course based assessment is less than it used to be. The Queensland system and the AQA model are based on many years of research, (Petch, 1967; Rooke and Hewitt, 1970; Smith, 1978; Maxwell, 2004; Cumming et al., 2004) both of which showed that constant meeting and discussion of coursework led to a raising of professional standards and an agreement of grades.

Wiliam (1998) has refers to this process as one of construct referencing. Neither norm referencing, nor criteria referencing in its strictest sense, construct referencing has elements of both and something else besides. It took a practice of impression marking, common amongst English teachers, one stage further. In essence, when English teachers award a grade to a piece of work, or a folder, they are using a construct of what they think that grade looks like, based on their previous encounters with work of a similar standard. Similar approaches, involving a combination of holistic judgements supported by more atomistic criteria, have been used in mathematics (Brown, 1992).

Sadler’s (1989) use of guild knowledge is also pertinent. He has argued that teachers use a sense of what it means to be very good at something and translate such knowledge, ‘guild knowledge’, into everything they mark. The final grade awarded mirrors how

---

3 For example, Graded Assessment in Mathematics (Brown, 1992) and the Association of Teachers of Mathematics (ATM) coursework-based examination (REF).
closely you approximate what it means to be good at that particular thing. Where formative assessment is involved, the teacher encourages the pupil, through peer marking and the like, to enter that guild.

The research context: the King’s Oxfordshire Summative Assessment Project

The project involved six English and six mathematics teachers in three secondary comprehensive schools together with researchers from King’s College London and Oxfordshire Local Authority. One of the teachers in each school pair was the Head of Department and the other an interested classroom teacher. The 3 sample schools were already well-versed in formative practices from the developmental work that continued throughout Oxfordshire schools following the King’s-Oxfordshire-Medway-Formative-Assessment (KMOFA) Project (1999–2001). The work on the KOSA project has focused on developing these teachers’ own summative assessment practices individually and in subject groups (within and between schools.) The current project ran over 30 months funded initially by the DfES, for the pilot phase, and in the main phase by the Nuffield Foundation.

Methodology
The project has used a design research methodology (Kelly, 2003) within a mixed methods approach, adopting and extending methods of data collection and analysis used in our previous research (Black, Harrison, Lee, Marshall, & Wiliam, 2003). A wide range of data has been collected including:

- evidence of summative assessments submitted for moderation;
- fieldnotes of all whole group and subject group project meetings;
- fieldnotes and audio recordings of intra-school standardisation and moderation meetings, as well as inter-school meetings preparing for these;
- classroom observations of summative assessment events;
- regular interviews with the 12 core teachers;
- focus group interviews with subject and school teacher groups examining similarities and differences to summative and formative assessment practices between subjects;
- teachers’ writing and reflections, in particular teachers reflective diaries.

The data sources on which this paper is based are the fieldnotes of meetings, interviews with teachers and evidence of summative assessments submitted for moderation. Data analysis has been conducted by the King’s team: coding transcripts and discussing emergent issues as in our previous work (Black et al., 2003; Hodgen & Marshall, 2005).

---

4 Because of promotions and personal reasons, the project involved eight English teachers and ten mathematics teachers. Most of the upheaval involving mathematics teachers was in one school. One mathematics teacher remained involved in the project after moving to another Oxfordshire school, whilst another transferred between two of the project schools.

5 Christine Harrison, Paul Black, Jeremy Hodgen, Bethan Marshall and Natasha Serret from King’s College London and Dorothy Kavanagh from Oxfordshire LA.
Quality and validity: seeking the “ideal” assessment

Notions of quality or what it meant to be “good at” English and mathematics formed the core of discussions with the teachers around the validity of judgements as to quality looked like in that subject. Throughout the project, the development of assessment tasks, the rationalisation and adjustment of judgements of students’ work, the interpretation of criteria and the evaluation of the tasks were often discussed with reference back to an underlying notion of quality. In other words, do these different elements address what it means to be good at mathematics or English? Both sets of teachers, English and mathematics, felt that the current assessment regimen did not adequately test what it meant to be good at the subject.

English teachers felt, particularly at KS3, that external tests lacked validity in that they did not capture what it meant good at the subject. During their interviews and project days, English teachers equated quality with, ‘insight’, ‘flair’, ‘confidence’ and an interest/enjoyment/engagement with language that allows for risk-taking and ‘subverting conventions’.

In mathematics, progression is very well articulated in terms of skills and procedures, but less well in terms of what the qualities the mathematics teachers described as ‘an intuitive understanding of number’, ‘understanding generalisation, ‘reasoning’ and being able to find the ‘structure of the problem and the mathematics that underlies it’.

Towards the end of the project, the English and mathematics teachers both reflected that the project had not brought about a significant change in this belief but had unwrapped what was latent. The project had ‘cemented’, ‘crystallised’, ‘legitimised’ their views. When asked, some reflected that they were able to articulate their views more clearly and quickly. For others, the project had encouraged them to add more detail, acknowledge a greater diversity in this view, especially in relation to looking at quality in terms of speaking and listening.

Tensions between the “ideal” and the reality of teacher assessment

What became increasingly apparent over the duration of the project was that the systems, tasks and methods that these teachers used for summative assessment often constrained their capacity to reinforce and communicate this view of quality in their daily classroom practice and interactions with students. Teachers explained this divide through a number of external and school pressures. Teachers interpreted recent government messages (scrapping GCSE Mathematics coursework) as a continued fostering of a culture of high stakes testing. This interpretation was reinforced by a wider public/parent perception of the subject, who ‘think that mathematics books need to have lots of ticks’ (Mathematics Teacher, June 2007), and what it means to be successful (getting good test scores). In their interactions with parents, teachers often felt that their ‘professional judgement’ was undermined as a result of this
You go to a parent’s evening. I will tell them what level they are working at and they ask, ‘but what will they get in the test?’ As if somehow this was far more reliable than my teacher assessment. (English teacher, May 2005)

For others, (particularly the mathematics teachers), this strengthened their argument that part of their professional responsibility is to prepare students for tests and rationalised why a good proportion of classroom time and summative assessment needs to be about preparing for and practising tests.

I am going to target those things. If you can do these things, it is going to radically improve the probability that you are going to get a C. Now that’s not necessarily making them a better mathematician, but that is saying if you practice these skills and learn to do these skills and apply them on that exam paper, you are more likely to get a C which will then enable, will open doors for you beyond Yr11 here. (Mathematics teacher, May 2005)

The real assessment, the proper assessment is the one they do in the exam. I think that this is a fairly common view, because it’s the one that goes on the certificate and it’s the one that makes a difference. It’s the one that gets them into college or gets them a job interview. It doesn’t matter if they have done brilliantly for five years of school. They do badly on one day and they get a D on a piece of paper and they are stuck. (Mathematics teacher, May 2005)

During the trial phase of the project, some of the mathematics staff started to recognise that this was leading to a teaching and learning experience divorced from their view of what it means to be good at the subject.

We are caught in a trap. Report to parents a level. Students focus on just doing tests so we have a level to report. We have missed the point of what we are trying to achieve. It is less and less with monitoring progress and more about filling in boxes. (Mathematics teacher, May 2005)

The teachers believed that the school leadership often exerted a pressure through the wider school assessment systems and policies put in place perhaps to secure greater consistency across curriculum subjects. In one school for example, the policy that teachers need to report to parents a termly ‘working at’ level and an end of year level created aggregation challenges for teachers wanting to use a greater diversity of evidence for a student. This appeared to make termly ‘in house’ tests more appealing.

At the beginning of the project, two of the three schools in English had done part of or all the QCA optional tests for their summer examinations. But by the end of the project all had abandoned theses tests. One school, however, was required to take the key stage 3 tests a year early, in year 8. All had some assessment done in controlled conditions. This meant one of the assignments was completed wholly in class and acted, therefore, as a good judge of the pupils’ independent capability.
Although there were considerable differences between the mathematics departments and all three department re-examined their tests as part of their project work, all three continued to use external tests alongside the teacher assessments for KOSAP. This was in part due to expectations from school management, but also as preparation for KS3 and GCSE and partly as a reality check on the teacher assessments. Hence, one of the schools used KS3 papers from previous years. Another school supplemented this with a regular (6 weekly) set of test developed to more closely align with KS3 tests. School management in the third had opted for an accelerated KS3, administering the KS3 national tests at Y8. The teachers also acknowledged some contradictions in their own beliefs saying that found it difficult to totally abandon belief in the validity of traditional tests. Some of the mathematics teachers also felt examinations to be motivating for students.

The decisions teachers made about the nature of the assessment tasks and, in particular, the consequent implications for moderation were influenced by their concerns for their department confidence and expertise especially for some mathematics project teachers. Tasks that attempted to communicate part of the view about what it means to be good at mathematics like ‘going down blind alleys’, ‘open-ended investigations’, ‘applying mathematics to real life problems’ tended to be less straightforward to mark and required greater interpretation and application of national curriculum level descriptors in order to assess student’s work.

The assessment process

The portfolio

A key aim of the project was to support teachers in reconsidering what a portfolio of assessment tasks might look like for a year 8 student in mathematics and English. This was already a common practice in English, but not in mathematics. Ideas for tasks were suggested, drafted, trialled, moderated and evaluated between May 2005 and July 2006 (as reported in Black et al., 2006). By September 2006, the project teachers were reaching some agreement. The portfolio needed to include a number of tasks that was reasonable enough to enable a student to demonstrate their overall attainment, yet manageable for marking, moderating and passing on from year to year (i.e., 6-9 pieces of work).

Collectively, the tasks had to offer students the opportunities to learn and demonstrate their ability in a variety of ways and contexts (i.e. small group discussion, role play, poster presentation, homework, paired, individual, whole class teacher) and reflect a balanced representation of all the curriculum attainment targets for that subject. Particular key areas within the attainment targets (for example, a lit crit piece) were identified and the number of tasks specified so that there was some commonality between portfolios across different schools.

In English, the Year 8 portfolio comprised seven tasks but they were assessed for nine items altogether: 2 writing (non fiction, fiction), 1 reading (lit crit), 1 controlled piece (reading and writing), 3 speaking and listening (group, role play, public speaking) –one of these Speaking and Listening piece will also be assessed for reading. In this way, at the end of Y8 a pupil will receive three marks/levels for each strand.
In Mathematics, the Y8 portfolio was designed to provide curriculum coverage and to allow both flexibility and comparability. Hence, it comprised six tasks: three common tasks, one in each of the curriculum strands (number & algebra, shape and space, data handling and all covering the using and applying mathematics strand) together with 3 non-common tasks. The pair of mathematics teachers from each school took responsibility for developing the task and assessment criteria for one of the three common tasks.

The English project teachers were more familiar with and had greater experience of a portfolio of assessment tasks. These teachers saw the project as refining their existing practice and more particularly formalising assessment requirements at Key stage 3- a phase that had received less attention compared with key stage 4. On the other hand, for the Mathematics project teachers used to regular testing, this had required them to invent new practice. Their experience was largely of GCSE coursework, but this is of much narrow range, has much greater specification than English and is generally disliked by mathematics teachers.. The teachers working in this area found the opportunity to explore/invent new tasks valuable, forcing them out of a habit of using off-the-shelf items.

**Teacher concerns and tensions about reliability**

Teacher concerns were greatest amongst the mathematics teachers. These concerns about reliability centred around two issues, the sampling and teacher input. Somewhat paradoxically, for a subject that includes the teaching of measurement and statistics, they felt any sampling procedures to be flawed because the sampling might fail to capture some critical aspect of a students’ learning. Ideally, all the mathematics teachers would have liked to be able to “look over all students’ shoulders all the time” (Meeting, 13 Oct 2005). One teacher’s comment was typical, “I want to know everything really” (Interview, Nov 2006). There was some contradiction here in that, when thinking of themselves as recipients of other teachers assessments, they largely wanted simply “a broad-brush level plus anything unusual” (Interview, Nov 2006). This concern with sampling led to several failed attempts to capture everything that was done or said by students in a lesson. Hence, they all identified class size as a significant obstacle to teacher assessment. Another strategy adopted by two of the mathematics departments was an attempt to assess all the work each student had done in class on a particular unit of work. When the teachers came to assess this work, the students’ classwork was very similar reflecting what had been taught rather than what had been learnt. This in turn led to a realisation amongst the mathematics teachers that (as with English) richer tasks were needed to give students opportunity to demonstrate what they could do. Hence, we concur with Crookes, Kane & Cohen’s (1996) argument that the solution to problems of reliability problem lie, at least in part, with validity.

The mathematics teachers’ perceived of a teacher input problem related to teacher bias and reflect what Gardner (2007) calls “the milieu of distrust of teacher assessments of students’ work” (p.20), although the teachers’ concerns were with unintentional rather than deliberate bias. These related to both other mathematics teachers’ strong belief in the validity of tests, a perception of poor teacher quality in mathematics and a view that
teachers would inadvertently advantage students by their input. For example, one teacher, a Head of Department, referred several times to other teachers conducting regular “sneaky” tests; she felt that, whilst the school’s policy was for comment-only marking, many of her colleagues found it difficult to ‘let go’ of tests. The concerns about bias also led to a focus on “controlled conditions”, specifying in fine detail the task, the time allocation and the teacher’s role. In particular, this included placing limits on the teacher’s input by drawing a common script for the mathematics activities in an attempt to teacher-proof the assessments. One school took the rather unusual step of changing teachers for the assessment activities, so that the teachers knew little about the students or their mathematical learning.

English teachers were not overly concerned with reliability, most of their queries being about validity. There is some truth in the teachers’ perception of poor teacher quality (e.g., Johnson-Wilder et al., 2003), but it is noteworthy that rather than seeing a focus on teacher assessment as a vehicle for professional learning and change, they perceived teacher quality as a barrier. This was despite their universal opinion that, for themselves, their KOSAP work had been a valuable learning experience.

Discussion

Guiding the project teachers’ choices and evaluations of what counts as a good task was the prevailing theme of what it means to be good at English and Mathematics and whether the tasks genuinely assessed qualities. On the whole, both the English and the Mathematics teachers were satisfied that the final selection of tasks did allow for this. One arising issue was that of differentiation. In English, some of the teachers reflected on how they would in the future ‘tweak’ the tasks by adjusting the activities or the resources so that the tasks were more effective at enabling students to demonstrate a wider range of attainment. Another challenge presented to teachers was where a task tried to assess two attainment targets. This crossover issue occurred in both English, where a task simultaneously assessed reading and writing and in using and applying mathematics set within the context of the content strands of Number and Algebra, Shape and Space or Data-handling. Maintaining a balance between the two targets and ensuring that one (e.g. writing) did not out weigh the other when making an overall judgement was identified as problematic but not impossible.

In both subjects, teachers expressed a keen interest in incorporating classroom talk in the portfolio. One major shortcoming of the current KS3 English test is that it contains no assessment of talk. The inclusion of Speaking and Listening in the portfolio, as an individual attainment target within the National Curriculum for English, created some advantages and disadvantages. The project teachers were able to draw from their experiences of GCSE coursework and used existing models to slot in Speaking and Listening assessment activities.
For some project mathematics teachers, managing the assessment of classroom talk, in their endeavour to collect evidence of everything to ensure greater reliability, created real organisational challenges, unless classroom talk could be formalised in some way, like an oral presentation. In both subjects, the project had heightened an awareness that ‘something needs to be done’ (English teacher) in this area and that more professional development is needed in how to instigate, generate, manage and capture classroom talk that acts as a tool for learning rather than a presentational device.

Many of the project teachers could already cite examples of the advantages that the portfolio offered in communicating to others, including parents and senior leadership teams. Task-in-portfolio work was something to talk with parents about as this work can have a cachet which normal class work lacks. Talk was perceived as a useful vehicle to assess weaker students, although actually assessing students’ talk posed manageability problems.

Teacher’s reflections over the National level descriptors as assessment criteria revealed a dilemma. For some English and Mathematics teachers, these criteria remained at the heart of the problem. They were too ‘vague’, ‘ambiguous’. One teacher (English) located this ambiguity in the lack of any coherent sense of progression where themes appear in level 3 and are then absent until level 7 or gradations in a skill (e.g. is beginning to…can use…is confident…) created too much room for interpretation. Other teachers (English and mathematics) worried that making the criteria too specific might constrain the learning potential of the task. These teachers were keen to retain a holistic impression of the student’s overall performance and to avoid the danger of the tasks becoming another ‘test that teachers teach to’.

Plagiarism and undue outside help were not seen as significant problems by any of the teachers: they felt that they knew their students well enough to spot when it happen and simply wished that the wider public had greater faith and confidence in their professional ability to do this. Some were in favour of pupils gaining credit if they have learnt more from legitimate influences from outside the school. They saw this as not dissimilar to what they do as teachers in the classroom, ‘… it won’t just jump out of them, we have to be skilful in how to get the best out of them’ (Mathematics teacher). This is a case where removal of advantages from the advantaged does not help the disadvantaged. Some teachers concluded that the inclusion of a ‘controlled conditions’ component in the portfolio was one step towards alleviating external concerns over the reliability of using tasks that were open to teacher/parent help and ‘cheating’.

---

6 An important distinction arises here for English. In English, Speaking and Listening is both a means and an object of assessment. In the paper, we use Speaking and Listening only to refer to English Attainment Target 1.
Conclusion: Emerging findings

As we have already noted, these findings are at this stage tentative and should be read with some caution. In addition, we have yet to analyse the data from moderation meetings. However certain important features appeared.

How important are subject differences?
As we expected, we found considerable differences between the two subjects. The English and mathematics teachers differed in their views about both reliability and validity, in the ways in which they used evidence to support their judgements and their attitudes to course work. In addition, the structure of the curriculum in the two subjects is different. This was particularly evident in the way that the Attainment Target 1 (Using and Applying Mathematics) is integrated within the content strands. These differences were reflected in the ways that two groups of teachers approached teacher assessment and to the value that they placed on it. Asked “cold” about the benefits of coursework, for example, this group of mathematics teachers would still be equivocal despite espousing the benefits whereas the English teachers were more supportive believing that coursework had a place. On the other hand, the differences within the subject groups appear to us to be as great. So, for example, as we have already discussed, both the subject groups differed in their views on the extent to which judgements should be made holistically or atomistically.

Transferring assessment expertise is not straightforward
Although we had expected the teachers to encounter difficulties with summative assessment, we were surprised by how difficult this was. As Ruthven argues, introducing new tools of any kind involves systemic change and is necessarily difficult. In addition, the teachers had some difficulty transferring their expertise between area. English teachers, whilst well versed, in making judgements in writing and reading, but encountered challenges in the assessment of speaking and listening. Most notably they felt that the criteria for assessing Speaking and Listening in the National Curriculum were inadequate and so they used GCSE criteria to make judgements. The mathematics teachers did have expertise in the context of GCSE coursework, but this was a much narrower and more tightly defined context than the mathematics curriculum in general.

Professional development: assessment literacy and values as well as skills
Gardner (2007) describes assessment expertise required by teachers as follows:

Assessment literacy would include knowledge of the types, methods and purposes of assessment, and an understanding of the reliability and validity, of responses, scores and grades, and so on, and their implications for learners. Assessment skills would in turn include competence in different assessment methods, question design, item writing, feedback, moderation, facilitation of self and peer assessment and so on. Assessment values would include an endorsement of the importance of consistency, impartiality and transparency in assessment practice. (p.20, our emphasis)
We find this a useful characterization. The English teachers increased in rigour as the project continued and aspired to literacy, skills and particularly values, wanting their assessment to be seen as consistent. The mathematics teachers had much stronger assessment skills (largely drawn from their experiences at GCSE) than either assessment literacy or values. Whilst there is certainly room for considerable improvement in the skills, we concur with Gardner that the main focus for professional learning should be on literacy and values. A particular issue in relation to assessment values was that the mathematics teachers appeared to judge teacher assessment in different terms to traditional assessment by tests. For example, whilst the mathematics teachers all felt that sampling was a critical issue for teacher assessment, they did not see this as a problem in relation to tests. English teachers tended to view each assessment, whether they were test or coursework, with scrutiny, although they tended to view tests as more likely to be invalid.

References


Manchester: Joint Matriculation Board.