‘In praise of educational research’: formative assessment

Paul Black and Dylan Wiliam

King’s College London

To appear in *British Educational Research Journal* v29 (2003)

Abstract

In this paper we trace the development of the King’s Formative Assessment Programme (KFAP) from its origins in diagnostic testing in the 1970s, through the graded assessment movement in the 1980s, to the present day. In doing so, we discuss the practical issues involved in reviewing research and outline the strategies we used to try to communicate our findings to as wide an audience as possible (including policy-makers and practitioners as well as academics). We then describe briefly how we worked with teachers to develop formative practice in classrooms, and discuss the impact that this has work has had on practice and policy. In the final section, we speculate about some of the reasons for this impact, and make suggestions for how the impact of educational research on policy and practice might be improved .

Introduction

It has long been recognised that assessment can support learning as well as measure it. Whilst it appears that Michael Scriven first used the term ‘formative evaluation’ in connection with the curriculum and teaching (Scriven, 1967), it was in Bloom, et al. (1971) that the term was the first used in its generally accepted current meaning. They defined *summative evaluation tests* as those tests given at the end of episodes of teaching (units, courses, etc) for the purpose of grading or certifying students, or for evaluating the effectiveness of a curriculum (p117). They contrasted these with “another type of evaluation which all who are involved—student, teacher, curriculum maker—would welcome because they find it so useful in helping them improve what they wish to do” (p117) which they termed ‘formative evaluation’.

From their earliest use it was clear that the terms ‘formative’ and summative’ applied not to the assessments themselves, but to the functions they served. However, the methods and questions of traditional summative tests might not be very useful for the purpose of the day-to-day guidance of learning. So the development of formative assessment depended on the development of new *tools*. To make optimum use of these teachers would also have to change their *classroom practices.* There would also be a need to align formative and summative work in new overall *systems,* so that teachers formative work would not be undermined by summative pressures, and indeed so that summative requirements might be better served by taking full advantage of improvements in teachers’ assessment work.

The rise and fall of formative assessment

In the 1970s and 1980s, the development of new *tools* was advanced by a series of research projects at Chelsea College (which merged with King’s College in 1985) which explored ways in which assessments might support learning. The Concepts in Secondary Mathematics and Science (CSMS) project investigated mathematical and

scientific reasoning in students through the use of tests that were intended to illuminate aspects of students' thinking, rather than just measure achievement.

In this venture, the science 'wing' of the project followed a broadly Piagetian approach, seeking to understand students' scientific reasoning in terms of stages of development (Shayer & Adey, 1981). This approach did not lead as directly to results applicable in normal teaching as did the more empirical approach of the mathematics team which focused on the diagnosis of errors in the concepts formed by secondary school students, and looked for ways to address them (Hart, 1981). Their subsequent projects sought to understand better the relationship between what was taught and what was learned (Johnson, 1989).

Such interest in the use of assessment to support learning was given added impetus by the recommendation of the Committee of Inquiry into the Teaching of Mathematics in Schools (1982) that a system of 'graded tests' be developed for pupils in secondary schools whose level of achievement was below that certificated in the current school-leaving examinations. Similar systems had been used to improve motivation and achievement in modern foreign languages for many years (Harrison, 1982).

The group at Chelsea College chose rather to aim at a system for all pupils, and with support from both the Nuffield Foundation and the Inner London Education Authority (ILEA), their Graded Assessment in Mathematics (GAIM) Project was established in 1983 It was one of five graded assessments schemes supported by the ILEA.

This development was more ambitious in attempting to establish a new *system*. In mathematics, English, and craft design and technology, the schemes set out to integrate the summative function with the formative. Information collected for formative purposes by teachers as part of the day-to-day classroom work would be aggregated at specific points in order to recognise the achievement of students formally. On leaving a school, the achievements to date would be 'cashed in' to secure a school-leaving certificate. This was allowed before 1988, but then the new criteria for the General Certificate of Secondary Education (GCSE) specified that, in mathematics, the assessments must include a written, end-of-course, examination which had to count for at least 50% of the available marks (Department of Education and Science & Welsh Office, 1985). The original developments in other subjects made more use of frequent formal tests, but were similarly constrained by the GCSE rules.

These rules were just one of three factors which undermined the graded assessment developments. The second was the introduction of a national curriculum for all schools in England and Wales in 1988. The National Curriculum Task Group on Assessment and Testing (TGAT 1988a, b) adopted the model of age-independent levels of achievement that had been used by the graded assessment schemes, but required a system of ten levels to cover the age range 5 to 16, arranged so that the average student could be expected to achieve one level every *two* years . This was too coarse-grained a system to be directly useful in classroom teaching: the graded assessment schemes needed twenty levels for the same age range. To add to the problem of re-calibrating the levels to the new national curriculum levels, a third problem became increasingly serious. Assuring comparability of awards, both between schools and with other traditionally-based awards, required costly administration. The combined effect of these three factors led, by 1995, to the abandonment of all the schemes.

Here began the decline of the development in formative assessment achieved in the 1970s and 80s. The graded assessment schemes had achieved remarkable success. In mathematics in particular they had given a clear indication of how formative and

summative functions might be integrated in a large-scale assessment system. They had also provided models of how reporting structures for the end-of-key-stage assessments might foster and support progression in learning, and also some ideas around how moderation of teachers' awards might operate in practice. Ironically, whilst these achievements had a strong influence on the TGAT's formulation of its recommendations, the new system that developed after their report served to undermine them.

The original national curriculum proposals had envisaged that much of the assessment at 7, 11 and 14 would:

> be done by teachers as an integral part of normal classroom work. But at the heart of the assessment process there will be nationally prescribed tests done by all pupils to supplement the individual teachers' assessments. Teachers will mark and administer these, but their marking – and their assessments overall – will be externally moderated. (Department of Education and Science & Welsh Office, 1987 p. 11)

Many of these suggestions were taken up in the TGAT's recommendations. In particular, the Task Group's first report pointed out that while fine-scale assessments which served a formative function could be aggregated to serve a summative function, the reverse was not true: assessments designed to serve a summative function could not be disaggregated to identify learning needs. The Task Group therefore recommended that the formative function should be paramount for key stages 1, 2 and 3. Many in the teaching profession welcomed the TGAT report, but the initial reception of the research community was quite hostile, as was that of the Prime Minister (Thatcher 1993 pp. 594-5)

Whilst the government accepted the recommendations of the Task Group (in a written parliamentary answer dated 7 June 1988), over the next ten years the key elements were removed, one by one, so that all that now remains in place is the idea of age-independent levels of achievement (Black, 1997; Wiliam, 2001). In particular, the idea that national curriculum assessment should support the formative purpose has been all but ignored. Daugherty (1995, p.78) points out that the words "teacher assessment" appeared on the agenda of the School Examinations and Assessment Council (the body responsible for the implementation of national curriculum assessment) only twice between 1988 and 1993.

Throughout the 1990s, the debate continued about how evidence from the external tests and teachers' judgements could be combined. In the end, it was resolved by requiring schools to publish data derived from teachers' judgements and those derived from the external tests, side-by-side. Whilst the government could claim that it was giving equal priority to the two components, in practice schools were not required to determine the final teacher judgements until *after* the test results had been received by the school. This created an incentive for teachers to match their own assessments to the results of the tests, rather than face criticism for having standards that were either too low or too strict.

So by 1995 nothing was left of the advances made in the previous decades. Government was lukewarm or uninterested in formative assessment: the *systems* to integrate it with the summative had gone, and the further development of *tools* was only weakly supported. There were some minor attempts in the 1990s by the government and its agencies to support formative assessment within the national curriculum assessment

system, including a project to promote the formative use of data from summative tests, but these efforts were little more than window dressing.

The debate over the relative weights to be applied to the results of external tests and teachers' judgements obscured the fact that both of these were summative assessments. While different teachers developed different methods of arriving at judgements about the levels achieved by their students (Gipps et al., 1995) it is clear that most of the record-keeping occasioned by the introduction of the national curriculum placed far more emphasis on supporting the summative function of assessment. At the same time, the work of Tunstall & Gipps (1996) showed that while some teachers did use assessment in support of learning, in many classrooms, it was clear that much classroom assessment did not support learning, and was often used more to socialise children than to improve achievement (Torrance & Pryor, 1998). Such studies were beginning to direct attention to the classroom *processes,* i.e. to the fine detail of the ways in which the day to day actions of teachers put formative principles into practices focused on learning.

Thus one sign of change was that within the educational research community, there was increasing concern that the potential of assessment to support learning was being ignored and there were continuing calls for the formative function of assessment to receive greater emphasis. The lead here was taken by BERA's Policy Task Group on Assessment. Their article (Harlen *et al*. 1992) and that of Torrance (1993) re-iterated the importance of the formative function of assessment, although there was disagreement about whether the formative and summative functions should be separated or combined (Black 1993b).

In 1997, as part of this effort to re-assert the importance of formative assessment, the BERA Policy Task Group on Assessment (with the support of the Nuffield Foundation) commissioned us to undertake a review of the research on formative assessment. They thereby initiated a new stage in the development of formative assessments.

Defining the field and reviewing the relevant research

Two substantial review articles, one by Natriello (1987) and the other by Crooks (1988) had addressed the field in which we were interested, and so our review of the research literature concentrated on articles published after 1987, which we assumed had been the cut-off point for the these earlier reviews. Natriello's review covered the full range of assessment purposes (which he classified as certification, selection, direction and motivation), while Crooks' review covered only formative assessment (which he termed 'classroom evaluation'). An indication of the difficulty of defining the field, and of searching the literature, is given by the fact that while the reviews by Natriello and Crooks cited 91 and 241 items respectively, only 9 references were cited in both.

Natriello's review used a model of the assessment cycle, beginning with purposes, and moving on to the setting of tasks, criteria and standards, evaluating performance and providing feedback and then discussed the impact of these processes on students. He stressed that the majority of the research he cited was largely irrelevant because of weak theorisation, which resulted in key distinctions (eg the quality and quantity of feedback) being conflated.

Crooks' paper had a narrower focus—the impact of evaluation practices on students. He concluded that the summative function of assessment has been too dominant and that more emphasis should be given to the potential of classroom assessments to assist learning. Most importantly, assessments must emphasise the skills, knowledge and attitudes regarded as most important, not just those that are easy to assess.

Our review also built on four other key reviews of research published since those by Natriello and Crooks—reviews by Bangert Drowns and the Kuliks into the effects of classroom testing (Bangert-Drowns et al., 1991a; Bangert-Drowns et al., 1991b; Kulik et al., 1990) and a review by one of us of research on summative and formative assessment in science education (Black 1993a).

To identify relevant literature we first searched the ERIC data-base. However, the lack of consensus on the appropriate key-words in this area meant that this process consistently failed to yield studies that we knew from our own reading to be relevant. Citation index searches on the existing reviews added additional studies, but still failed to identify others. Study of articles cited in the studies we had identified yielded some further studies. However, it was clear overall that we needed a different approach if we were to identify all of the many relevant studies. So we resorted to a manual search through every issue, from 1987 to 1998, of 76 of the most likely journals. As we read through these journals manually, rather that relying on keywords, our notion of what was relevant was continually expanding.

This process generated 681 publications that appeared, at first sight to be relevant. By reading abstracts, and in some cases full papers, this number was reduced to 250 publications that were read in full, and coded with our own keywords. Keywords were then grouped to form sections of the review. In synthesising the studies allocated to each section, we rejected the use of meta-analysis. In the first place, many of the studies did not provide sufficient details to calculate standardised effect sizes, which is the standard procedure for combining results from different studies in meta-analysis (Glass at al., 1981). Second, the use of standardised effect sizes in published studies over-estimates the size of actual effects. This is because the low statistical power of most experiments in the social sciences (Cohen, 1988) means that many experiments generate results that fail to reach the threshold for statistical significance, and go unpublished, so that those that are published are not a representative sample of all results actually found (Harlow et al., 1997). Thirdly, and most importantly in our view, given the relatively weak theorisation of the field, we felt it was not appropriate simply to specify inclusion criteria and then include only studies that satisfied them. Instead, we assembled the review through a process of 'best evidence synthesis' (Slavin, 1990). Such an approach is inevitably somewhat subjective, and is in no sense replicable– other authors, even given the same 250 focal studies, would have produced a different review. Instead, our aim was to produce an account of the field that was authentic, faithful, and convincing. However, it is worth noting that none of the six short commentaries by experts in this field, which were published in the same journal issue as our review, disagreed with our findings.

We believe that this story of the development of our review makes several important points about educational research. In the first place, reviewing research is much more difficult in any social science than it is in (say) the physical sciences because the complexity of the field precludes any simple universal system of key words. Any automated research process is bound to result in a systematic under-representation of the body of research because it cannot hope to identify all the relevant studies.

Second, synthesising research cannot be an objective process. While review protocols such as those being used by the EPPI-Centre for the evaluation of individual research studies are helpful in identifying strengths and weaknesses in those studies, the significance attached to each study, and the way that general conclusions are drawn by weighing sometimes conflicting findings, will inevitably remain subjective.

Thirdly, it seems to us important to point out that reviewing research is not merely a derivative form of scholarship. Reviews such as those by Natriello and Crooks can serve to reconceptualise, to organise, and to focus research. Because our definition of 'relevance' expanded as we went along, we too had to find ways of organising a widening field of research,

and were forced to make new conceptual links just in order to be able to relate the various research findings into as coherent a picture as possible. This was one reason why our review generated a momentum for work in this field that would be difficult to create in any other way.

One feature of our review was that most of it was concerned with such issues as students' perceptions, peer- and self-assessment, and the role of feedback in a pedagogy focused on learning. Thus it helped to take the emphasis in formative assessment studies away from *systems,* with its emphasis on the formative-summative interface, and re-locate it on classroom *processes*.

Disseminating the findings

Experience in producing a review of mathematics education research for teachers (Askew & Wiliam, 1995) had taught one of us that it is impossible to satisfy, in the same document, the demands of the academic community for rigour and the demands for accessibility by practitioners. Discussions with primary teachers showed that even when that draft review had been written in the most straightforward language we could manage, they still found it unsuitable. Extensive re-drafting eventually made it accessible, but this work took as long the original collection and writing, and it then turned out that the academic community found it of little interest.

Such experience convinced us that we had to produce different outputs for different audiences. For other academics, we produced a 30,000-word journal article (Black & Wiliam, 1998a), which, together with short responses from invited commentators from around the world, formed the whole of a special issue of the journal *Assessment in Education*. As well as detailing our findings, we tried to lay out as clearly as possible how we had constructed the review so that, while we would not necessarily expect different authors to reach identical conclusions, we hoped that the process which we followed was verifiable and could be repeated.

To make the findings accessible to practitioners and policy makers, we produced a twenty-one page booklet in A5 format entitled *Inside the black box* (Black & Wiliam, 1998b). We also produced a slightly revised version for the international audience (Black & Wiliam, 1998c).

In *Inside the black box*, we very briefly outlined the approach we had taken in our review and summarised its conclusions. However, we also felt it essential to explore implications for policy and practice. In so doing we inevitably, at some points, went beyond the evidence, relying on our experience of many years work in the field. If we had restricted ourselves to only those policy implications that followed logically and inevitably from the research evidence, we would have been able to say very little. Whilst the policy implications we drew were not *determined* by the research base, they were, we felt the most reasonable interpretation of the findings (for more on reasonableness as a criterion, see below). The title itself was significant, for it pointed to our main policy plea – that teachers' work in the classroom was the key to raising standards and that systems of external testing should be re-structured to ensure that this work was supported rather than undermined by them. The *process* now claimed priority in the *system*.

In order to secure the widest possible impact of our work, we planned the booklet's launch with members of the King's College London's External Relations Department. We held a launch conference on 5 February 1998, hosted by the Nuffield Foundation, who had supported the writing of the review, and on the previous day had held a series

of briefings for journalists of the national daily newspapers and the specialist educational press.

The result was that almost every single national daily newspaper carried some reference to the work. While some of the reports were either inaccurate or highly politicised, much of the coverage was broadly accurate, if somewhat selective.

In the five years since its launch, *Inside the black box has* sold over 30,000 copies and data from the Authors' Licensing and Collecting Society (ALCS) suggests that at least another 25,000 copies have been made in schools.

Inside the black box did not try to lay out what formative assessment would look like in practice. Indeed, it was clear that this was neither advisable nor possible:

> Thus the improvement of formative assessment cannot be a simple matter. There is no 'quick fix' that can be added to existing practice with promise of rapid reward. On the contrary, if the substantial rewards of which the evidence holds out promise are to be secured, this will only come about if each teacher finds his or her own ways of incorporating the lessons and ideas that are set out above into her or his own patterns of classroom work. This can only happen relatively slowly, and through sustained programmes of professional development and support. (pp 15)

Putting it into practice

The three research reviews offered strong evidence that improving the quality of formative assessment would raise standards of achievement in each country in which it had been studied. Furthermore, the consistency of the effects across ages, subjects and countries meant that even although most of the studies reviewed had been conducted abroad, these findings could be expected to generalise to the United Kingdom. For us, the question was therefore not "Does it work?" but "How do we get it to happen?"

In mid-1998 the Nuffield Foundation agreed to support a two-year project to involve 24 teachers in six schools in two LEAs (Oxfordshire and Medway) in exploring how formative assessment might be put into practice. So began the King's-Medway-Oxfordshire Formative Assessment Project (KMOFAP).

One of the key assumptions of the project was that if the promise of formative assessment was to be realised, traditional research designs—in which teachers are 'told' what to do by researchers—would not be appropriate. We argued that a process of supported development was an essential next step:

> In such a process, the teachers in their classrooms will be working out the answers to many of the practical questions that the evidence presented here cannot answer, and reformulating the issues, perhaps in relation to fundamental insights, and certainly in terms that can make sense to their peers in ordinary classrooms. (p.16 in Black & Wiliam, 1998b)

We had chosen to work with Oxfordshire and Medway because we knew that their officers were interested in formative assessment and would be able to support our work locally. We asked each authority to select three secondary schools and, after discussion with members of the research team, the six so chosen agreed to be involved. At this stage we were not concerned to find 'typical' schools, but schools that could provide 'existence proofs' of good practice, and would produce the 'living examples' alluded to earlier for use in further dissemination.

We decided to start with mathematics and science because these were subjects where we felt there were clear messages from the research and also where we had expertise in the subject specific details that we thought essential in practical development. The choice of teachers, two in mathematics and two in science in each school, was left to the schools: a variety of methods were used, so that there was a considerable range of expertise and experience amongst the 24 teachers selected.

In the following year we augmented the project with one additional mathematics and one additional science teacher, and also began working with two English teachers from each school.

Our 'intervention' with these teachers had two main components:

> a series of nine one-day in-service (INSET) sessions over a period of 18 months, during which teachers were introduced to our view of the principles underlying formative assessment, and were given the opportunity to develop their own plans;
>
> visits to the schools, during which the teachers would be observed teaching by project staff, and have an opportunity to discuss their ideas and their practice; feedback from the visits helped us to attune the INSET sessions to the developing thinking and practice of the teachers.

The key feature of the INSET sessions was the development of action plans. Since we were aware from other studies that effective implementation of formative assessment requires teachers to re-negotiate the 'learning contract' that they had evolved with their students (Brousseau, 1984; Perrenoud, 1991), we decided that implementing formative assessment would best be done at the beginning of a new school year. For the first six months of the project (January 1999 to July 1999), therefore, we encouraged the teachers to experiment with some of the strategies and techniques suggested by the research, such as rich questioning, comment-only marking, sharing criteria with learners, and student peer- and self-assessment. Each teacher was then asked to draw up an action plan of the practices they wished to develop and to identify a single focal class with whom these strategies would be introduced in September 1999. Details of these plans can be found in Black *et al*. (2003).

Our intervention did not impose a model of 'good formative assessment' on teachers, but rather supported them in developing their own professional practice. Since each teacher was free to decide which class to experiment with, we could not impose a standard experimental design—we could not standardise the outcome measures, nor could we rely on having the same 'input' measures for each class. In order to secure quantitative evidence, we therefore used an approach to the analysis that we have termed 'local design', making use of whatever data were available within the school in the normal course of events. In most cases, these were the results on the national curriculum tests or GCSE but in some cases we also made use of scores from school assessments. Each teacher consulted with us to identify a focal variable (i.e. dependent variable or 'output') and in most cases, we also had reference variables (i.e. independent variables or 'inputs). We then set up, for each experimental class, the best possible control class in the school. In some cases, this was a parallel class taught by the same teacher (either in the same or previous years); in others, it was a parallel class taught by a different teacher. Failing that, we used a non-parallel class taught by the same or different teacher. We also made use of national norms where these were available. In most cases, we were able to condition the focal variable on measures of prior achievement or general ability. By dividing the differences between the mean scores of control group and experimental groups by the pooled standard deviation, we were able

to derive a standardised effect size (Glass et al., 1981) for each class. The median effect size was 0.27 standard deviations, and a jack-knife procedure (Mosteller & Tukey, 1977) yielded a point estimate of the mean effect size as 0.32, with a 95% confidence interval of [0.16, 0.48]. Of course we cannot be sure that it was the increased emphasis on formative assessment that was responsible for this improvement in students' scores, but this does seem the most reasonable interpretation (see discussion of 'reasonableness' below). For further details of the experimental results, see Wiliam, Lee, Harrison & Black (2003).

Outcomes

The quantitative evidence that formative assessment does raise standards of achievement on national curriculum tests and GCSE examinations is important in showing that innovations that worked in research studies in other countries could also be effective in typical UK classrooms. Part of the reason that formative assessment works appears to be an increase in students' 'mindfulness' (Bangert-Drowns et al., 1991a), and while this has been shown to increase long-term retention (Nuthall & Alton-Lee, 1995), it also depends to a certain extent on the kind of knowledge that is being assessed. More will be gained from formative feedback where a test calls for the mindfulness that it helps to develop. Thus it is significant here that almost all the high-stakes assessments in this country require constructed responses (as opposed to multiple choice) and often assess higher-order skills.

However, other outcomes from the project are at least as important. Through our work with teachers, we have come to understand more clearly how the task of applying research into practice is much more than a simple process of 'translating' the findings of researchers into the classroom. The teachers in our project were engaged in a process of knowledge creation, albeit of a distinct kind, and possibly relevant only in the settings in which they work (see Hargreaves, 1999).

As the teachers explored the relevance of formative assessment for their own practice, they transformed ideas from other teachers into new ideas, strategies and techniques, and these were in turn communicated to other teachers, creating a 'snowball' effect. Also, as we have introduced more and more teachers to these ideas, we have become better at communicating the key ideas. A case in point is that we have each been asked several times by teachers, "What makes for good feedback?"—a question to which, at first, we had no good answer. Over the course of two or three years, we have evolved a simple answer — good feedback causes thinking.

Ever since the publication of *Inside the black box* we have been producing a series of articles in journals aimed at practitioners that showed how the ideas of formative assessment could be used in practice. We are also publishing research papers arising from the project. These have included accounts of our collaborative work with teachers, of the processes of teacher change, of the quantitative evidence of learning gains, and of a theoretical framework for classroom formative assessment. (see the project's web-site at http://www.kcl.ac.uk/ depsta/education/KAL/ASSESSMENT.html for details of the programme's publications).

In addition, since the launch of *Inside the Black Box* in February 1998, members of the research team at King's have spoken to groups of teachers and policy-makers on over 400 occasions, suggesting that we have addressed well over 20,000 people directly about our work.

Following the success of *Inside the black box*, we decided to use the same format for communicating the results of the KMOFAP project to teachers. The resulting booklet, *Working inside the black box*, sold 15,000 copies in the six months after its launch in July 2003.

Of course working as intensively as we did with 24 teachers could not possibly impact more than a small fraction of teachers, so we have also given considerable thought to how the work could be 'scaled up'. We are working with other LEAs (notably Hampshire) to develop local expertise, in both formative assessment and in strategies for dissemination. In Scotland, formative assessment has become an important component of the Scottish Executive's strategy for schools and members of the project team, including the project teachers, have provided support and advice.

This work has also lead to further research work. We are partners, with the Universities of Cambridge, Reading, and the Open University, in the *Learning How to Learn: in classrooms, schools and networks* project, which is part of the ESRC's Teaching and Learning Research Programme (TLRP). This project aims to promote and understand change in classrooms, but also is investigating how these changes are supported or inhibited by factors at the level of the whole school, and of networks of schools. We are also working with Stanford University in California on a similar project, which again combines detailed work with small numbers of teachers with a focus on how small-scale changes can be scaled up. Assessment for learning has also become one of the two key foci (along with thinking skills) of the government's Key Stage 3 Strategy for the foundation subjects.

Reflections

Critiques of educational research typically claim that it is poorly focused, fails to generate reliable and generalisable findings, is inaccessible to a non-academic audience and lacks interpretation for policy-makers and practitioners (Hillage *et al*. 1998). While some of these criticisms are undoubtedly applicable to some studies, we believe that this characterisation of the field as a whole is unfair.

We do not believe that all educational research should be useful, for two reasons. The first is that, just as most research in the humanities is not conducted because it is useful, we believe that there should be scope for some research in education to be absolutely uninterested in considerations of use. The second reason is that it is impossible to state, with any certainty, which research will be useful in the future.

Having said this, we believe strongly that the majority of research in education should be undertaken with a view to improving educational provision— research in what Stokes (1997) calls "Pasteur's quadrant". And although we do not yet know everything about 'what works' in teaching, we believe that there is a substantial consensus on the kinds of classrooms that promote the best learning. What we know much less about is how to get this to happen.

Policy-makers appear to want large-scale research conducted to the highest standards of analytic rationality, whose findings are also relevant to policy. However, it appears that these two goals are, in fact, incompatible. Researching how teachers take on research, adapt it, and make it their own is much more difficult than researching the effects of different curricula, of class sizes, or of the contribution of classroom assistants. While we do not know as much as we would like to know about effective professional development, if we adopt 'the balance of probabilities' rather than 'beyond reasonable

doubt' as our burden of proof, then educational research has much to say. When policy without evidence meets development with some evidence, development should prevail.

In terms of our own work, perhaps the most puzzling issue arising from this story is why our work has had the impact that it has. Our review did add to the weight of evidence in support of the utility of formative assessment, but did not substantially alter the conclusions reached by Crooks and Natriello ten years earlier. It could be, of course, that the current interest in formative assessment, and its policy impact, is nothing to do with our work, but that it takes ten years or so for such findings to filter through to policy, or that the additional studies identified by us in some way tipped the balance to make the findings more credible. All of this seems unlikely. It therefore appears that something that we did, not just in undertaking the review, but also in the way it was disseminated, has had a profound impact on how this research has fed into policy and practice. Of course identifying which elements have been most important in this impact is impossible, but it does seem appropriate to speculate on the factors that have contributed to it.

The first factor is that although most of the studies cited in our review emanated from overseas, the review was originated and published in this country. This provided a degree of local 'ownership' that is likely to have attracted attention to the research.

The second is that although we tried to adhere closely to the traditional standards of scholarship in the social sciences when conducting and writing our review, we did not do so when exploring the policy implications in *Inside the black box*. While the standards of evidence we adopted in conducting the review might be characterised as those of 'academic rationality', the standard within *Inside the black box* is much closer to that of 'reasonableness' advocated by Stephen Toulmin for social enquiry (Toulmin, 2001). In some respects, *Inside the black box* represents our opinions and prejudices as much as anything else, although we would like to think that these are supported by evidence, and are consistent with the 50 years of experience of working in this field that we have between us. It is also important to note in this regard that the success of *Inside the black box* has been as much due to its rhetorical force as to the evidence that underpins it. This will certainly make many academics uneasy—for it appears to blur the line between fact and value, but as Flyvbjerg (2001) argues, social enquiry has failed precisely because it has focused on analytic rationality rather than value-rationality (see also Wiliam, 2003).

The third factor we believe has been significant is the steps we have taken to publicise our work. As noted above, this has included writing different kinds of articles for different audiences, and has also involved working with media-relations experts to secure maximum press coverage. Again, these are not activities that are traditionally associated with academic research, but they are, we believe, crucial if research is to impact on policy.

A fourth factor that appears to have been important is the credibility that we brought as researchers to the process. In their project diaries, several of the teachers commented that it was our espousal of these ideas, as much as the ideas themselves, that persuaded them to engage with the project: where educational research is concerned, the facts do not necessarily speak for themselves.

A fifth factor is that the ideas had an intrinsic acceptability to the teachers. We were talking about improving learning in the classroom, which was central to the professional identities of the teachers, as opposed to bureaucratic measures such as target-setting.

Linked to this factor is our choice to concentrate on the classroom processes and to live with the external constraints operating at the formative-summative interface: the failed attempts to change the *system* in the 80s and 90s were set aside. Whilst it might have been merely prudent to not try again to tilt at windmills, the more fundamental strength was that it was at the level chosen, that of the core of learning, that formative work stakes its claim for attention. Furthermore, given that any change has to work out in teachers' practical action, this is where reform should always have started. The evidence of learning gains, from the literature review and from our project, restates and reinforces the claim for priority of formative work that the TGAT tried in vain to establish. The debate about how policy should secure optimum synergy between teachers' formative, teachers' summative, and external assessments is unresolved, but the terms and the balance of the arguments have been shifted.

The final, and perhaps most important fact in all this is that in our development model, we attended to both the content and the process of teacher development (Reeves *et al*. 2001). We attended to the proces*s* of professional development through an acknowledgement that teachers need time, freedom, and support from colleagues, in order to reflect critically upon and to develop their practice (Lee, 2000), whilst offering also practical strategies and techniques about how to begin the process. By themselves, however, these are not enough. Teachers also need concrete ideas about the directions in which they can productively take their practice, and thus there is a need for work on the professional development of teachers to pay specific attention to subject-specific dimensions of teacher learning (Wilson & Berne, 1999).

One might ask of our work whether it drew, or could have drawn, strength from the psychology of learning. For the practical demand for good *tools*, e.g. a good question to explore ideas about the concept of momentum, the help could only come from studies in mathematics and science education. We have shown earlier how the historical origins of our work contributed to such studies. The fact that the King's team had strong subject backgrounds and could draw on earlier work on such *tools* was essential, However, it remains the case that for most school subjects rich sources for the appropriate *tools* are lacking.

At a more general level however, it could be seen that the practical activities developed did implement principles of learning that are prominent in the psychology literature. Examples are the constructivist principle that learning action must start from the learner's existing knowledge, the need for active and responsible involvement of the learner, the need to establish in the classroom a community of subject discourse, and the value of developing meta-cognition (see Black et al. 2003). Our orientation in developing the activities in line with such principles was in part intuitive, but became explicit when the teachers asked us to give a seminar on learning theory at one of the group's INSETs. We judge now that to have started with such principles when we had little idea of how to implement them in practice might have done little to motivate teachers.

One feature that emerges from considering the history of this work is that its success seems far from inevitable, but is rather the result of a series of contingencies. What would have happened if our review article had not been commissioned, or if we had finished work when the review was complete (which is all we had been commissioned to do)? What if we had failed to secure funding for our project, or had failed to find teachers willing to work with us, or if we worked in a department forced to make cuts in staffing as a result of a research assessment exercise? The continuity of staffing seems to us to be particularly important. We do not think that it is a co-incidence that the work here builds on a 25-year tradition of work on diagnostic and formative assessment

within a single institution, and one in which the turnover of staff has been extremely low. While some of the expertise gathered over this time can be made explicit, much of it cannot, and we believe that our collective implicit knowledge has been at least as important to our work as published research findings.

This history paints a picture of educational research not as a steady building up of knowledge towards some objective truth about teaching and learning, but rather as a trajectory buffeted by combinations of factors. Calling this "the mangle of practice" Pickering (1995) shows that neither the view of scientific knowledge as a series of truths waiting to be discovered, nor the view that scientific knowledge is whatever scientists do, are adequate descriptions. Instead, he suggests, scientists build ideas about how the world is, which are then 'mangled' in their contact with the objective physical world. When questions become too difficult, scientists change the questions to ones that are tractable. In social sciences, the mangle is even more complex, in that what it is possible to research, and what it is good to research keeps changing.

From this perspective, all our activities—the development of theoretical resources, work with teachers, rhetorical campaigns to convince policy makers and teachers of the utility of formative assessment, even co-opting the print and broadcast media to our purpose—makes a kind of sense.

Educational research can and does make a difference, but it will succeed only if we recognise its messy, contingent, fragile nature. Some policy makers believe that supporting educational research is crazy, but surely the real madness is to carry on what we have been doing, and yet to expect different outcomes.

References

Askew, M. & Wiliam, D. (1995). *Recent research in mathematics education 5-16*. London, UK: Her Majesty's Stationery Office.

Bangert-Drowns, R. L.; Kulik C-L, C.; Kulik, J. A. & Morgan, M. T. (1991a). The instructional effect of feedback in test-like events. *Review of Educational Research,* **61 2)**, 213 - 238.

Bangert-Drowns, R. L.; Kulik, J. A. & Kulik, C.-L. C. (1991b). Effects of frequent classroom testing. *Journal of Educational Research,* **85**(2), 89-99.

Black, P. (1997). Whatever happened to TGAT? In C. Cullingford (Ed.) *Assessment vs. evaluation* (pp. 24-50). London, UK: Cassell.

Black, P.; Harrison, C.; Lee, C.; Marshall, B. & Wiliam, D. (2003). *Assessment for learning: putting it into practice*. Buckingham, UK: Open University Press.

Black, P. J. (1993a). Formative and summative assessment by teachers. *Studies in Science Education,* **21**(1), 49-97.

Black, P.J. (1993b). Assessment policy and public confidence : Comments on the BERA Policy Task Group's article ' Assessment and the improvement of education'. *The Curriculum Journal*. **4**(3), 421-427.

Black, P. J. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles Policy and Practice,* **5**(1), 7-73.

Black, P. J. & Wiliam, D. (1998b). *Inside the black box: raising standards through classroom assessment*. London, UK: King's College London School of Education.

Black, P. J. & Wiliam, D. (1998c). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan,* **80**(2), 139-148.

Bloom, B. S.; Hastings, J. T. & Madaus, G. F. (Eds.). (1971). *Handbook on the formative and summative evaluation of student learning*. New York, NY: McGraw-Hill.

Brousseau, G. (1984). The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics. In H.-G. Steiner (Ed.) *Theory of mathematics education: ICME 5 topic area and miniconference* (pp. 110-119). Bielefeld, Germany: Institut für Didaktik der Mathematik der Universität Bielefeld.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Committee of Inquiry into the Teaching of Mathematics in Schools (1982). *Report: mathematics counts*. London, UK: Her Majesty's Stationery Office.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research,* **58**(4), 438-481.

Daugherty, R. (1995). *National curriculum assessment: a review of policy 1987-1994*. London, UK: Falmer Press.

Department of Education and Science & Welsh Office (1985). *General Certificate of Secondary Education: the national criteria*. London, UK: Her Majesty's Stationery Office.

Department of Education and Science & Welsh Office (1987). *The National Curriculum 5-16: a consultation document*. London, UK: Department of Education and Science.

Flyvbjerg, B. (2001). *Making social science matter: why social inquiry fails and how it can succeed again*. Cambridge, UK: Cambridge University Press.

Gipps, C. V.; Brown, M. L.; McCallum, E. & McAlister, S. (1995). *Intuition or evidence? Teachers and the national assessment of seven year olds*. Buckingham, UK: Open University Press.

Glass, G. V.; McGaw, B. & Smith, M. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Hargreaves, D.H. (1999). The knowledge creating school. *British Journal of Educational Studies*, **47** (2), 122-144.

Harlen, W., Gipps, C., Broadfoot, P. and Nuttall, D. (1992). Assessment and the improvement of education. *The Curriculum Journal,* **3** (3), 215 - 230.

Harlow, L. L.; Mulaik, S. A. & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Harrison, A. (1982). *Review of graded tests*. London, UK: Methuen.

Hart, K. M. (Ed.) (1981). *Children's understanding of mathematics: 11-16*. London, UK: John Murray.

Hillage, J.; Pearson, R.; Anderson, A. & Tamkin, P. (1998). *Excellence in research on schools*. London, UK: Department for Education and Employment.

Johnson, D. C. (Ed.) (1989). *Children's mathematical frameworks 8-13: a study of classroom teaching*. Windsor, UK: NFER-Nelson.

Kulik, C.-L. C.; Kulik, J. A. & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: a meta-analysis. *Review of Educational Research,* **60**(2), 265-299.

Lee, C. (2000, September) *The King's Medway Oxford Formative Assessment Project: studying changes in the practice of two teachers.* Paper presented at Symposium entitled 'Getting Inside the Black Box : Formative Assessment in Practice' at the British Educational Research Association 26th annual conference held at Cardiff University. London, UK: King's College London School of Education.

Mosteller, F. W. & Tukey, J. W. (1977). *Data analysis and regression: a second course in statistics*. Reading, MA: Addison-Wesley.

National Curriculum Task Group on Assessment and Testing (1988a). *A report*. London, UK: Department of Education and Science.

National Curriculum Task Group on Assessment and Testing (1988b). *Three supplementary reports*. London: Department of Education and Science.

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist,* **22**(2), 155-175.

Nuthall, G. & Alton-Lee, A. (1995). Assessing classroom learning: how students use their knowledge and experience to answer classroom achievement test questions in science and social studies. *American Educational Research Journal,* **32**(1), 185-223.

Perrenoud, P. (1991). Towards a pragmatic approach to formative evaluation. In P. Weston (Ed.) *Assessment of pupil achievement* (pp. 79-101). Amsterdam, Netherlands: Swets & Zeitlinger.

Pickering, A. (1995). *The mangle of practice: time, agency, and science*. Chicago, IL: University of Chicago Press.

Reeves, J.; McCall, J. & MacGilchrist, B. (2001). Change leadership: planning, conceptualization and perception. In J. MacBeath & P. Mortimore (Eds.), *Improving school effectiveness* (pp. 122-137). Buckingham, UK: Open University Press.

Scriven, M. (1967). *The methodology of evaluation*. Washington, DC: American Educational Research Association.

Shayer, M. & Adey, P. S. (1981). *Towards a science of science teaching: cognitive development and curriculum demand*. London, UK: Heinemann Educational Books.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: a best evidence synthesis. *Review of Educational Research,* **60**(3), 471-499.

Stokes, D. E. (1997). *Pasteur's quadrant: basic science and technological innovation*. Washington, DC: Brookings Institution Press.

Thatcher, M. (1993). *The Downing Street Years*, London: Harper Collins

Torrance, H. (1993). Formative assessment: some theoretical problems and empirical questions. *Cambridge Journal of Education,* **23**(3), 333-343.

Torrance, H. & Pryor, J. (1998). *Investigating formative assessment*. Buckingham, UK: Open University Press.

Toulmin, S. (2001). *Return to reason*. Cambridge, MA: Harvard University Press.

Tunstall, P. & Gipps, C. (1996). Teacher Feedback to Young Children in Formative Assessment: a typology. *British Educational Research Journal,* **22**(4), 389-404.

Wiliam, D. (2001). *Level best? Levels of attainment in national curriculum assessment*. London, UK: Association of Teachers and Lecturers.

Wiliam, D. (2003). The impact of educational research on mathematics education. In A. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Second International Handbook of Mathematics Education* (pp. 469-488). Dordrecht, Netherlands: Kluwer Academic Publishers.

Wiliam, D.; Lee, C.; Harrison, C. & Black, P. (2003). Teachers developing assessment for learning: impact on student achievement. *Submitted for publication*.

Wilson, S. M. & Berne, J. (1999). Teacher learning and the acquisition of professional knowledge: an examination of research on contemporary professional development. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (pp. 173-209). Washington, DC: American Educational Research Association.

Acknowledgements

We would like to acknowledge the initiative of the Assessment Policy Task Group of the British Educational Research Association (now known as the Assessment Reform Group) who gave the initial impetus and support for our research review. We are grateful to the Nuffield Foundation, who funded the original review, and the first phase of our project. We are also grateful to Professor Myron Atkin and his colleagues in Stanford University, who secured funding from the US National Science Foundation (NSF Grant REC-9909370), for the last phase. Finally, we are indebted to the Medway and Oxfordshire local education authorities, their six schools and above all to their thirty-six teachers who took on the central and risky task of turning our ideas into practical working knowledge.