LINKING RESEARCH AND PRACTICE: KNOWLEDGE TRANSFER OR KNOWLEDGE CREATION?

Dylan Wiliam
King's College London
dylan.wiliam@kcl.ac.uk

In D. S. Mewborn, P. Sztajn, D. Y. White, H. G. Wiegel, R. L. Bryant, & K. Nooney (Eds.), Proceedings of *Twenty-fourth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* conference, vol 1 (pp. 51-69). Columbus, OH: Educational Resources Information Centre (ERIC) Clearinghouse on Science, Mathematics and Environmental Education, 2002.

In this paper, it is argued that the failure of educational research to impact on practice stems from a failure to understand the nature of expertise in teaching, and that traditional models of knowledge *transfer* can only be effective for those at a relatively limited level of competence. Instead, it is suggested that teachers need to be involved, collaboratively, with researchers in a joint process of knowledge *creation*. In the King's-Medway-Oxfordshire Formative Assessment Project (KMOFAP) a group of 24 secondary school teachers (grades 6 to 12) of mathematics and science were supported in developing action plans of how they wanted to develop their classroom assessment practice with a single class, through a series of four day-long workshops, and by observations of their teaching. Comparison with local controls (established on a case-by-case basis for each teacher) on curriculum-based tests, showed an average effect size of +0.32.

Introduction

Although the amount of money spent on educational research in most countries over the last hundred or so years has only been a tiny fraction of the expenditure on education (ranging from one-third of a percent to one percent in most developed countries in recent years) a large amount of money has certainly been spent on educational research, and yet its impact is very hard to discern.

The failure of educational research to have any real impact on educational practice in general, and on mathematics education in particular, has been lamented for many years. Today, there are, it seems, two broad strands of criticism of research in education. The first is that educational research is unnecessary. This manifests itself either in a belief that expert practitioners already know 'what works' in mathematics classrooms and so novice mathematics teachers can learn all they need to know by watching experienced teachers, or that pedagogical practice will always be weak, and that the solution lies in prescribing curricula and teaching methods in 'teacher-proof' schemes.

The second strand regards educational research as necessary, but of poor quality. Too often, it is said, educational research produces results that are ambiguous or contradictory, perhaps best summed up by Robert F Kennedy's furious reaction to the ambiguous evaluation of the impact of additional money provided for the education of socioeconomically disadvantaged students: "Do you mean that you spent a billion dollars and you don't know whether they can read or not?" (Lagemann, 2000 p202). On those few occasions when research does produce unambiguous results, these are generally felt to tell us what we already knew. The lament continues: If educational researchers could only agree how to go about research properly, then educational research could join the elite club of 'hard' sciences producing reliable knowledge (these people have in the past rather unkindly been described as suffering from 'physics envy').

In this paper, I want to suggest is that by its very nature, by the nature of the things it studies, educational research cannot produce 'reliable knowledge' in the sense that Physics—the paradigm case—has done. Modern Physics may be complex, but its successes have been secured because the things it studies, complex as they are, are actually rather simple compared to educational processes. In education, the pursuit of Grand Unified Theories that provide reliable descriptions of what to do in every situation is doomed to fail. Instead, I want to argue that educational research should be about the pursuit of 'practical wisdom' about how educational processes can be improved, and a necessary corollary of this will be that educational research cannot be done 'on' teachers, but only with them, and that this should be not a process of getting teachers to do what we want them to do (co-operation) but of creating knowledge, with teachers (collaboration).

In doing so, I hope to show that such a shift does not entail a down-grading of educational research to a pseudo science, but that, as was first argued by Aristotle the pursuit of practical wisdom is actually a higher-order goal than the pursuit of pure knowledge. The remainder of the paper then goes on to describe how these ideas about collaborative research were put into practice with a group of teachers in the King's-Medway-Oxfordshire Formative Assessment Project (KMOFAP).

The (troubling) history of educational research

The history of educational research can be viewed as a search for disciplinary foundations. At the beginning of the last century, educational research, to the extent that it existed at all, was either historical or an aspect of philosophy. One of the earliest attempts to use empirical methods in educational research was the

'School Survey' movement in the United States. Beginning around 1910 this movement sought to gather 'objective evidence' about factors influencing the educational progress of school students. However, due to the sheer diversity of the United States education system, with over 100,000 school districts each free to determine its own education policy, there was little agreement about the purpose or scope of education, and meaningful comparisons of educational outputs was almost impossible.

In *An Elusive Science* (whose subtitle is the title of this section) Ellen Condliffe Lagemann (2000) shows that the search for ways of producing high-quality research in education has been, in effect, a search for secure disciplinary foundations for the production of reliable knowledge. At first, philosophy and history provided those foundations but, around the turn of the century, these were supplanted by psychology, which dominates to the present day, although since the 1970s sociology and social anthropology have also been influential.

Lagemann argues that the failure of educational research to deliver what has been wanted has arisen from three main causes—the isolation and low status of educational research in the academy; its tendency to focus too narrowly on particular aspects of education rather than looking at education systems; and the weak governance and regulation of educational research. These three causes are of course intimately entwined.

In the United States, teaching had been regarded as 'women's work' since early in the nineteenth century, so that educational research was accorded low status by association. Lagemann also points out that being an applied subject served to marginalia education within the academic mainstream. No doubt partly in an attempt to raise its status, educational research attempted to emulate the hard sciences through the quantification of educational processes, which of course entailed focusing on those aspects of education that could be easily quantified. And while most teachers were female, most school supervisors and district administrators were male, so that the emerging field of educational research emphasised educational administration almost from the outset.

This lack of agreement about not just how to undertake educational research, but also what should be researched continued to plague attempts to establish 'what works' in education over the next half-century, but Lagemann's history closes with an ironic twist. In the final quarter of the last century, educational research finally began to get on the right track with two key realisations. Firstly, the complexity of educational settings requires that insights from all of the 'foundation disciplines' (and not just one) are required to make progress in educational improvement. Secondly, it slowly became clear that centre-periphery models of dissemination are simply ineffective in education. The result was a blossoming of multi-disciplinary research, involving teachers in real innovation and improvement. However, at the same time, it seems that the politicians gave up on educational research, and by 1991, federal funding for educational research in the USA stood at approximately one-third the level provided in 1971.

While Lagemann's analysis is persuasive, I want to suggest here that the goal of educational research as a science is not just elusive, but impossible. This is in part a philosophical claim, but it is also in part an empirical claim. The phenomena that are studied in educational research are, in the first instance, far more complex than those that are studied by the 'hard' sciences—just imagine trying to set boundary conditions for the initial state of a typical mathematics classroom. However, in addition, it is important to realise the autonomy that individual students bring to lessons is not a problem with which physical sciences have to grapple. Bars of iron do not behave differently because someone has been testing them. Or more precisely, while bars of iron may behave differently depending on how they have been treated in the past (ie whether they have been annealed or subject to repeated stress and strain), we know what kinds of treatments matter, and we know how to find out in advance how the bar will perform under tests. Even those who believe that there is no such thing as free will, and that all human behaviour at time T1 is actually determined by the state of the system at T0 have to concede that it is too difficult to specify the starting conditions precisely enough to determine the outcome. Chaos theory, and, at a smaller scale, Heisenberg's uncertainty principle, renders Laplace's dream of being able to predict all behaviour from initial conditions a non-starter.

Expertise

There are also reasons to suspect that the nature of expertise in teaching is not reducible to the kinds of recipes used in the physical sciences. Flyvbjerg (2001) reports an experiment that was conducted on a group of paramedics (Klein & Klein, 1981). Six short video extracts of a person administering cardiopulmonary resuscitation (CPR) were shown to experienced paramedics, students being trained as paramedics, and people who taught life-saving techniques. They were then asked which of the six they would choose to resuscitate them if they needed CPR. Five of the six video extracts were of inexperienced trainees just learning CPR while the sixth was a highly experienced paramedic. Of the experienced paramedics, 90% chose the experienced paramedic, while only 50% of the students did so. However, only 30% of the instructors chose the experienced paramedic.

Flyvbjerg argues that we can understand this apparently paradoxical result by considering the five levels of expertise in learning proposed by Dreyfus and Dreyfus (1986). At the *novice* level, the individual is guided by rules that are applied irrespective of context. The novice teacher tends to try to apply the same sets of rules

to all the classes they teach. The *advanced beginner* begins to take situationally-specific factors into account, and personal experience is often relied on more than context-independent rules. However, as experience accumulates, the number of recognisable elements or 'chunks' increases, and threatens to overwhelm the individual. For example, the need for the school teacher to attend to the learning needs of her or his students, controlling the behaviour of some, while also trying to make sure that they interact as much with female and male students, can lead to a feeling of 'plate spinning'—dashing from one imperative to the next to try to attend to all. The *competent performer* is characterised by performance in which conflicting priorities are resolved through the use of strategies, usually derived from conscious problem-solving behaviour. In contrast, the *proficient performer* acts quickly and intuitively often doing the 'right thing' without conscious awareness. In this context, it is important to realise that 'intuition' is used here not as some irrational prejudice, but rather as the result of the sedimentation and synthesis of vast amounts of experience. Finally, in the *expert*, the ability to act quickly and intuitively, in a range of contexts and settings, is unified into a 'feeling' of the right thing to do. The use of an emotive term—'feeling'—here is not coincidental. Experts 'feel' the best course of action, not just with their mind, but in their whole body. Expertise is therefore not the culmination of rationality, but rather transcends it. Expertise involves going beyond what can be done through rationality, not irrational, but meta-rational (ie beyond rationality).

Therefore, Flyvbjerg argues, it appears that the paramedic trainers identified the trained paramedics less successfully because they looked for paramedics who followed the rules that they themselves taught. In other words, they were looking for those at the level of competent performers, rather than proficient performers or experts. If we accept that the classification proposed by Dreyfus and Dreyfus also applies to teaching, then it seems likely that the failure of educational research to impact on educational practice stems from a similar limitation.

The kinds of prescriptions given by educational research to practice have been in the form of generalised principles, that may often, even usually, be right, but are sometimes just plain wrong. The expert can see that a particular recipe is inappropriate in some circumstances, although because their response is intuitive, may not be able to discern the reason why. What gets learnt by the practitioner is that the findings of educational research are not a valid guide to action.

But research findings also run foul of the opposite problem—that of insufficient specificity. Many teachers complain that the findings from research produce only bland platitudes, that are insufficiently contextualised to be used in guiding action in practice. Put simply, research findings underdetermine action.

Knowledge transfer and knowledge creation

If we accept that the prime (although not the only) purpose of educational research is the improvement of educational processes, then research findings must be taken up by teachers and incorporated into their practice. There are other ways that educational research might influence practice—through the improvement of textbooks for example—but without some change in those who use them, innovations are unlikely to have much effect. In the past, this process has been called dissemination, and is now more often called knowledge transfer—both interesting metaphors, suggesting that all that needs to be done is to inform practitioners about the latest findings and they will be used. If expertise transcends rationality, as I have argued above, however, then the process of knowledge transfer cannot be one of providing instructions to novices, advanced beginners, or competent performers in the hope that they will get better. Rather what is needed is an acknowledgement that what teachers do in 'taking on' research is not a more or less passive adoption of some good ideas from someone else, but an active process of knowledge creation:

> Teachers will not take up attractive sounding ideas, albeit based on extensive research, if these are presented as general principles which leave entirely to them the task of translating them into everyday practice—their classroom lives are too busy and too fragile for this to be possible for all but an outstanding few. What they need is a variety of living examples of implementation, by teachers with whom they can identify and from whom they can both derive conviction and confidence that they can do better, and see concrete examples of what doing better means in practice. (Black & Wiliam, 1998b p15)

The different ways in which knowledge is transferred and created within organisations has been studied by Nonaka and Takeuchi (1995) who have proposed a simple framework for knowledge creation in organisations. In their model, there are four modes of knowledge conversion, depending on whether knowledge is converted to or from implicit or explicit knowledge (see figure 1). .

Figure 1 about here

The traditional kind of knowledge conversion practised by educational researchers is what Nonaka and Takeuchi call *combination*. Knowledge in an explicit form is converted to more knowledge in explicit form. At the other extreme, *socialisation* is their name for the process by which new practitioners become enculturated into new practices which are not known explicitly to those who are learning, nor to those from whom they are learning. Tacit knowledge becomes explicit knowledge through a process of *externalisation*, and explicit knowledge becomes implicit by *internalisation*. A learning cycle can then be set up in which knowledge is created, transformed, and circulated around an organisation. Through *learning by doing* systemic knowledge becomes operationalised, which can then be *shared* with other practitioners. In dialogue with others conceptual knowledge is built up, which is then combined with that of others through networking. It was this knowledge cycle that we attempted to implement in the King's-Medway-Oxfordshire Formative Assessment Project (KMOFAP) funded initially by the Nuffield Foundation (as the *Developing Classroom Practice in Formative Assessment* project) and subsequently by the United States National Science Foundation through their support of our partnership with the Stanford CAPITAL project (NSF Grant REC-9909370)

Collaborating with teachers: the KMOFAP project

Reviews of research by Natriello (1987) and Crooks (1988) and more recently by Black and Wiliam (1998a) had demonstrated that substantial learning gains are possible when teachers integrate assessment with classroom instruction. However, it is also clear from these reviews, and from other studies (see Black and Atkin 1996) that achieving this is by no means straightforward. As Black and Wiliam (1998b) point out:

> 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. This does not weaken the message here—indeed, it should be a sign of its authenticity, for lasting and fundamental improvements in teaching and learning can only happen in this way (pp15-16).

The challenge for us, then, was how could teachers be supported in incorporating formative assessment (or assessment for learning as it is sometimes called) into their classroom practice, not as a 'bolt on' series of tactics, but integrated into planning and teaching?

The research strategy

The central tenet of the research 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, for all the reasons discussed above—would not be appropriate. We therefore decided that we had to work in a genuinely collaborative way with a small group of teachers, beginning in the bottom left-hand corner of Nonaka and Takeuchi's model, by sharing with them our understanding of the research literature. We then invited the teachers to explore some of these ideas for themselves, by trying them out in their own classrooms (internalisation). At first, they were hesitant. Although we told them that we did not have a clear plan for what they should do, the teachers did not believe this. They seemed to believe that we were operating with a perverted model of discovery learning in which we knew full well what we wanted the teachers to do, but wouldn't tell them, because we wanted the teachers 'to discover it for themselves'. However, after a while, it became clear that there was no prescribed model of effective classroom action, and each teacher would need to find their own way of implementing the general principles of high-quality classroom assessment in their own classrooms. We then planned that they would share their experiences with other teachers in the group, and develop a common way of thinking about classroom assessment (socialisation). Through extended dialogue, we hoped that they could then develop a common language of description (externalisation) thus yielding findings that could be made explicit, so beginning another cycle (combination).

The sample

We began by selecting two school districts where we knew there was support from the authority for attempting to develop formative assessment, and, just as importantly, where there was an individual officer who could act as a link between the research team and the schools, thus providing a local contact for ad hoc support for the teachers. In this regard, we are very grateful to Sue Swaffield from Medway and Dorothy Kavanagh from Oxfordshire who, on behalf of their authorities, helped to create and nurture our links with the schools. Their involvement in both planning and delivering the formal inservice sessions, and their support 'on the ground' were invaluable, and it is certain that the project would not have been as successful without their contributions.

Having identified the two districts, we asked each district to select three schools that they felt would be suitable participants in the project. We were very clear that we were not looking for 'representative' or typical schools. From our experiences in curriculum development—for example in graded assessment (Brown, 1988)—we were aware that development is very different from implementation. What we needed were schools that had already begun to think about developing 'assessment for learning', so that with these teachers we could begin to produce the 'living examples' alluded to earlier to use in further dissemination.

Each district identified three schools that were interested in exploring further the possibility of their involvement, and the project directors visited each school with the officer from the school district to discuss the project with the principal and other members of the senior management team. All six schools identified agreed to be involved. Brief details of the six schools are shown in table 1.

In our original proposal to the Nuffield Foundation, we had proposed to work only with mathematics and science teachers, partly because of our greater expertise in these subjects, but also because we believed that the implications for assessment for learning were clearer in these areas. In order to avoid the possible dangers of isolation, our design called for two mathematics and two science teachers at each school to be involved.

Table 1 about here

The choice of teachers was left to the school, and a variety of methods was used. In some schools, the principals nominated a faculty chair together with a relatively inexperienced teacher. In other schools, teachers appeared to be selected because, in the words of one head, "they could do with a bit of inset [professional development]". In the event while our schools were not designed to be representative, there was a considerable range of expertise and experience amongst the 24 teachers selected.

The intervention

The intervention had two main components:

a) a series of half-day and one-day professional development days, during which teachers would be introduced to our view of the principles underlying formative assessment, and have a chance to develop their own plans;

b) visits to the schools, during which the teachers would be observed teaching by project staff, and have an opportunity to discuss their ideas, and how they could be put into practice more effectively

The pattern of professional development sessions is shown in table 2 (subsequent events were held as part of the NSF-funded work on the CAPITAL project, but the data reported here relate to the original project, from January 1999 to August 2000.

Table 2 about here

The key feature of the 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 'didactic contract' (Brousseau, 1984) that they had evolved with their students, 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, 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, and later to refine, an action plan specifying which aspects of formative assessment they wished to develop in their practice and to identify a focal class with whom these strategies would be introduced in September 1999. Although there was no inherent structure in these plans, the teachers being free to explore whatever they wished, we did find that they could be organised under the broad headings shown in table 3. In all the 24 teachers included a total of 102 activities in their action plans—an average of just over four each—and while there were a small number of cases of teachers of the same subject at the same school adopting common plans, there was no other clustering of teachers discernible. Inevitably the clear phases suggested by Nonaka and Takeuchi's model became increasingly blurred over the course of the project, with discussion frequently involving all four modes. While it has not been useful for analysis of the data arising from the project, nevertheless, we believe that the model provided a useful framework for shaping our initial interventions.

Table 3 about here

Most of the teachers' plans contained reference to two or three important areas in their teaching where they were seeking to increase their use of classroom assessment generally followed by details of strategies

that would be used to make this happen. In almost all cases the plan was given in some detail, although many teachers used phrases whose meanings differed from teacher to teacher (even within the same school).

Practically every plan contained a reference to focusing on or improving the teacher's own questioning techniques although not all of these gave details of the particular way in which they were going to do this (for example using more open questions, allowing students more time to think of answers or starting the lesson with a focal question). Others were less precise (for example using more sustained questioning of individuals, or improving questioning techniques in general). Some teachers mentioned planning and recording their questions. Many teachers also mentioned involving students more in setting questions (for homework, or for each other in class)**.** Some teachers also saw existing national tests as a source of good questions.

Using comment-only grading was specifically mentioned by nearly half the teachers, although many foresaw problems with this, given school policies on grading. Four teachers planned to bring forward end-of-topic tests thus providing time for remediation.

Sharing the objectives of lessons or topics was mentioned by most of the teachers, through a variety of techniques (using a question that the students should be able to answer at the end of the lesson , stating the objectives clearly at the start of the lesson, getting the students to round up the lesson with what they had learned). About half the plans included references to helping the students understand the grading criteria used for investigative or exploratory work, generally using exemplars from students from previous years. Exemplar material was mentioned in other contexts such as having work on display and asking students to correct work using a set of criteria provided by the teacher.

Almost all the teachers mentioned some form of self-assessment in their plans, ranging from using red, amber or green 'traffic lights' [stop lights] to indicate the student's perception of the extent to which a topic or lesson had been understood, to strategies that encouraged self-assessment via targets which placed responsibility on students (eg one of these twenty answers is wrong: find it and fix it!). Traffic lights (or some equivalent) were seen in about half of the plans and in practically all cases their use was combined with strategies to follow up the cases where the students signalled incomplete understanding.

Several teachers mentioned their conviction that group work provided important re-inforcement for students, as well as providing the teacher with insights they into their students' understanding of the work.

The choices of activities by the different teachers also shown no particular pattern, as the multidimensional scaling (Schiffman, Reynolds, & Young, 1981) of these data in figure 2 shows.

Figure 2 about here

The other component of the intervention, the visits to the schools, provided an opportunity for project staff to discuss with the teachers what they were doing, and how this related to their efforts to put their action plans into practice. The interactions were not directive, but more like a holding up of a mirror to the teachers. Since project staff were frequently seen as 'experts' in either mathematics or science education, there was a tendency sometimes for teachers to invest questions from a member of the project team with a particular significance, and for this reason, these discussions were often more effective when science teachers were observed by mathematics specialists, and vice-versa.

We aimed for each teacher to be observed six times over the school year from September 1999 to July 2000, although releasing teachers to discuss their lessons either before or afterwards was occasionally a problem (and schools that had guaranteed teacher release for this purpose at the beginning of the project were sometimes unable to provide for this).

Research design

Given the nature of the intervention, which was designed to build on the professionalism of teachers (rather than imposing a model of 'good formative assessment' on them), we felt that to utilise a traditional research design on the teachers would have been inconsistent. Furthermore, it would have been impractical. Since each teacher was free to choose which class would be the focus for this work, there was no possibility of standardising either the 'input' or 'output' variables. For this reason, the collection of empirical quantitative data on the size of effects was based on an approach which we have termed 'local design'. Drawing more on interpretivist than positivist paradigms, we sought to make use of whatever assessment instruments would have been administered by the school in the normal course of events. In many cases, these were state-mandated assessments such as the national tests for 14-year-olds or grades achieved in the national school leaving examinations (the General Certificate of Secondary Education or GCSE). In other cases we made use of scores from school assessments (particularly in science, where 'modular' approaches meant that scores on end-of-module tests were available). For each teacher we therefore had a focal variable (ie dependent variable or 'output') and in all but a few cases, we also had reference variables (ie independent

variables or 'inputs). In order to be able to interpret the outcomes we discussed the local circumstances in their school with each teacher and set up the best possible control group consistent with not disrupting the work of the school. In some cases this was a parallel class taught by the same teacher in previous years (and in one case in the same year). In other cases, we used a parallel class taught by a different teacher and, failing that, a non-parallel class taught by the same or different teacher. We also made use of national norms where these were available. In almost all cases, we were able to condition the focal variable on one or more reference variables, although in some cases the reference variables were measures of aptitude (eg NFER's Cognitive Abilities Test) while in others they were measures of achievement (eg end-of-year 8 tests).

In order to be able to compare the results, raw differences between experimental and control groups were standardised by dividing by the pooled standard deviation of the experimental and control scores.

Results

Table 4 provides the results achieved by the 19 teachers for whom controls were available and the standardised effect sizes are summarised in stem-and-leaf form in figure 3. As can be seen, the majority of effect sizes are around 0.2 to 0.3, with a median value of 0.27 and a mean of 0.34. Since the effect sizes were not normally distributed, the jack-knife procedure recommended by Mosteller and Tukey (1977) was used which provides an estimate of the true mean as 0.32 and a 95% confidence interval of the true effect size as (0.16, 0.48).

Figure 3 about here

In order to examine the relationship between a teacher's practice and the effect sizes, we classified teachers into one of four groups, according to their use of formative assessment strategies in their classrooms:

Experts	Formative assessment strategies embedded in and integrated with practice
Competent performers	Teachers who were successful with one or two key strategies, but having routinised these, were looking for other ways to augment their practice
Advanced beginners	Teachers who were successful with one or two key strategies, and had restricted themselves to these
Novices	Teachers who had attempted strategies , but had not embedded any strategies into their practice.

These characterisations had emerged from our observations of each teacher's practice, and were based on their use of key strategies during the main period of the project. Independent classification of the 24 teachers by the two main researchers produced identical classification for all but two teachers, and these were resolved after discussion. The effect sizes by teacher type are shown in table 5. Although there is no obvious trend in terms of average effect size, as one moves from less to more expert teachers, the interquartile range of effect sizes reduces, indicating further support for the attribution of the effects to the quality of classroom assessment.

Table 5 about here

A comparison of the effects by different forms of control in the form of side-by-side stem-and-leaf diagrams (figure 4) shows that no significant difference in effect sizes for the different form of controls is apparent.

Figure 4 about here

There was no difference in the mean effect size for groups of different ages, although it is worth pointing out that the year 11 (grade 10) focal groups all had positive effect sizes. There was no systematic variation in effect size by track. Analysis by subject shows that all the negative effect sizes were found for the mathematics groups, although the median effect sizes for the mathematics and science groups were almost identical.

Discussion

By its very nature, the quantitative evidence provided here is difficult to interpret. The controls are not equally robust. In some cases, we have comparisons with the same teacher teaching a parallel class in previous years, which, in terms of the main question (ie has the intervention had an effect?) is probably the

best form of control. In other cases, we have comparisons only with a different teacher teaching a parallel class, so it could be that in some cases a positive effect indicates only that the teacher participating in the project is a better teacher than the control. In other cases, the control is another class (and sometimes a parallel class) taught by the same teacher, and while there are examples of positive effect sizes here (in the case of Robert, for example) it is also reasonable to assume that the observed size of such effects will be attenuated by what we have termed 'uncontrolled dissemination'. In some cases, the only controls available were the classes of different teachers teaching non-parallel classes, and given the prevalence of ability-grouping in mathematics and science, and its effect on achievement (see Wiliam & Bartholomew, 2001) disentangling the effect of our interventions from contextual factors is quite impossible. However, given the fact that the outcome variables were either national tests and examinations, or assessments put in place by the school, rather than devised by the teacher, we have some confidence that these measures have some validity in terms of what the teachers were trying to achieve, and, more importantly, that teachers do not have to choose between teaching well and getting good results.

However, the improvements in the achievements of students is not the only (nor perhaps the most important) outcome of this project. While a small number of our teachers did view involvement in the project as a short-term commitment, after which they would return to teaching 'normally', for the vast majority of our teachers, involvement in the project has not just spread to all their classes, but has fundamentally altered their views of themselves as professionals (Black, Harrison, Lee, Marshall and Wiliam, 2002). They not only enjoy their teaching more, but have become ambassadors spreading the message to other teachers.

Conclusion

In this paper, I have argued that by trying to emulate the 'hard sciences' educational research has taken a wrong turn. Expertise in teaching does not consist of more and more complex explicit schemes for determining action (as is the case for example, in quantum physics), but is, rather, beyond rationality. Expertise is the ability to 'feel' what is the right thing to do, not after long deliberation, but immediately, and intuitively and this intuition is not instinctive, but is the result of the sedimentation of vast numbers of examples of experience. The role of the researcher in supporting the development of such expertise is not to attempt to distil expertise down to its essence, but to encourage its development in others. We cannot 'bottle' this for widespread distribution, but we can support communities of teachers by highlighting profitable directions in which they might develop their practice. At the end of the project, we are left with a final irony. In allowing the teachers to choose what they developed in their practice (so that each teacher was, in effect, engaged in a unique experiment) we have given up the ability to say what worked. We know that the process in which the teachers were engaged was productive, but we cannot say which elements worked, and which did not. In allowing the teachers to create their knowledge, we have given up the ability, as researchers, to make our own particular knowledge claims. So be it.

References

Black, P. J. & Atkin, J. M. (Eds.). (1996). Changing the subject: innovations in science, mathematics and technology education. London, UK: Routledge.

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 (also in Phi Delta Kappan **80**(2) 139-148).

Black, P.; Harrison, C.; Lee, C.; Marshall, B. & Wiliam, D. (2002). *Working inside the black box: assessment for learning in the classroom*. London, UK: King's College London Department of Education and Professional Studies.

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.

Brown, M. L. (Ed.) (1988). Graded Assessment in Mathematics development pack: teacher's handbook. Basingstoke, UK: Macmillan.

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

Dreyfus, H. & Dreyfus, S. (1986). *Mind over machine: the power of human intuition and expertise in the era of the computer*. New York, NY: Free Press.

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

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

Klein, H. A. & Klein, G. A. (1981) *Perceptive/cognitive analysis of proficient cardio-pulmonary resuscitation (CPR) performance*. Paper presented at Annual meeting of the Midwestern Psychological Association held at Chicago, IL.

Lagemann, E. C. (2000). *An elusive science: the troubling history of education research*. Chicago, IL: Chicago University Press.

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

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

Nonaka, I. & Takeuchi, H. (1995). *The knowledge-creating company: how Japanese companies create the dynamics of innovation*. New York, NY: Oxford University Press.

Schiffman, S. S.; Reynolds, M. L. & Young, F. W. (1981). Introduction to multidimensional scaling: theory, methods, and applications. New York, NY: Academic Press.

Wiliam, D. & Bartholomew, H. (2001, September) The influence of ability-grouping practices on student achievement in mathematics. Paper presented at British Educational Research Association 27th annual conference held at University of Leeds. London, UK: King's College London School of Education.

School	Abbreviation	Description
Brownfields	BF	Boys
Century Island	CI	Mixed
Cornbury Estate	CE	Mixed
Riverside	RS	Mixed
Two Bishops	TB	Mixed
Waterford	WF	Girls

Table 1: the six schools involved in the project

Inset	held		format	focus
A	February	1999	whole-day, London	introduction
B	May	1999	whole-day, London	developing action plans
C	June	1999	whole-day, London	reviewing and revising action plans
	September	1999	half-day, district-based	reviewing and revising action plans
D	November	1999	whole-day, London	sharing experiences, refining action plans, planning dissemination
E	January	2000	whole-day, London	research methods, dissemination, optional sessions including theories of learning
F	April	2000	whole-day, London	integrating learning goals with target setting and planning, writing personal diaries
G	June	2000	whole-day, London	action plans and school dissemination plans, data analysis 'while you wait'

Table 2: pattern of professional development events

Category	Activity	Code	Frequency
Questioning	Teacher questioning	TQ	11
	Pupils writing questions	PWQ	8
	Existing assessment: pre-tests	EAPT	4
	Pupils asking questions	PAQ	4
Feedback	Comment-only marking	COM	6
	Existing assessment: re-timing	EART	4
	Group work: test review	GWTR	4
Sharing criteria with learners	Course work: marking criteria	CWMC	5
	Course work: examples	CWEG	4
	Start of lesson: making aim clear	SoLMAC	4
	Start of lesson: setting targets	SoLST	1
	End of lesson: teacher's review	EoLTR	1
	End of lesson: pupils' review	EoLPR	4
	Group work: explanation	GWExp	2
	Involving classroom assessment	ICA	2
Self-assessment	Self-assessment: traffic lights	SATL	11
	Self-assessment: targets	SAT	5
	Group work: test review	GWTS	6
	Self-assessment: other	SAO	7
	Pupil peer-assessment	PPA	5
	Group work: revision	GWRev	1
General	Including parents	IncP	1
	Posters	Post	1
	Presentations	Pres	1
Total			102

Table 3: frequencies of activities in the action plans of 24 teachers

School	Subj	Teacher	Yr	Set	n	Focal variable	Reference variables	Control group	n	SD	Raw effect	*d*	p
BF	M	Iwan	7	1	25	SE7	C7	D	95	17.54	+6.63	**+0.38**	0.0299
BF	M	Iwan	9	1	27	KS3	C7, S8	D	94	33.84	12.25	**+0.36**	0.0081
BF	M	Lily	7	3	25	SE7	C7	D	95	14.96	-5.22	-0.35	0.1434
BF	S	Rose	7	5	8	SE7	SB7	D	25	24.80	38.44	**+1.55**	0.0001
BF	S	Peter											
CE	M	Belinda	8	1	21	SE8	KS2	P	26	10.61	2.76	**+0.26**	0.3604
CE	M	Angela	9	3	23	KS3	KS2	D	21	15.93	19.12	**+1.20**	0.0001
CE	S	Sian	8	-	26	SE8	SE7	P	169	0.889	0.342	**+0.38**	0.0113
CE	S	Carl	8	-	27	SE8	SE7	P	169	0.911	0.417	**+0.46**	0.0018
CI	S	Derek	9	2	27	KS3	C7	D	56	0.666	0.183	+0.27	0.1984
CI	S	Philip	9	1		KS3	C7	P	56	0.695	0.169	+0.24	0.2305
CI	M	Greg	9	4	24	KS3	SE7	P	20	0.0379	-0.025	-0.07	0.8045
CI	M	Eva	9	1	29	KS3	SE7	P	28	0.4916	-0.127	-0.26	0.3997
RS	M	Nancy	8	1	32	KS3	C7	P*	34	38.7	-12	**-0.31**	0.0019
RS	M	Nancy	8	1	32	KS3	C7	S	30	27.8	+32	**+1.15**	0.0001
RS	M	Nancy	9	1	34	KS3	KS2	N	-	0.50	0.13	+0.26	0.0669
RS	M	Patrick	9	1	30	KS3	KS2	N	-	0.58	0.38	**+0.66**	0.0001
RS	M	Lisa											
RS	S	Jerry	8	2									
RS	S	Tom	8	2	32	SE8	-	P	34	43.38	+10.02	+0.23	0.3852
TB	S	James	11	1	32	GCSE	-	S	32	0.879	0.255	+0.29	0.2628
TB	S	James	11	1	32	GCSE	-	P	32	1.013	0.375	+0.38	0.1038
TB	S	Robert	9	-	30	KS3	SE8	I	56	15.33	2.95	+0.19	0.1438
TB	M	Steve	11	2	32	GCSE	KS3	P	31	0.941	0.380	+0.40	0.1093
TB	M	Steve	11	4	24	GCSE	KS3	D	87	1.48	0.222	+0.15	0.2849
TB	M	Kerry	11	4	23	GCSE	KS3	D	87	1.54	0.309	+0.20	0.1348
TB	M	Kerry	11	1	32	GCSE	KS3	D	82	1.95	0.4786	**+0.25**	0.0276
WF	M	Gwen	9	2	23	KS3	-	L	24	0.462	0.158	**+0.34**	0.2469
WF	M	Alice											
WF	S	Susan											
WF	S	Kieron											

Key

Focal variables

KS3 Key stage 3 tests
SBn School-produced test at beginning of year n
SEn School-produced test at end of year n

Reference variables

Cn CAT scores in year n
SEn School produced tests at end of year n

Controls

I Parallel track taught by same teacher in same year
S Similar track taught by same teacher in previous year
P Parallel track taught by different teacher in same year
L Similar track taught by different teacher in previous year
D Non-parallel track taught by different teacher in same year
N National norms
* Non-representative control

Table 4: experimental results for the 24 teachers

Group	Count	Median	Interquartile range
Experts	7	0.25	0.07
Moving pioneers	10	0.31	0.25
Static pioneers	2	1.38	0.35
Trialers	6	0.15	0.64

Table 5: Effect sizes classified by teachers' use of formative assessment strategies

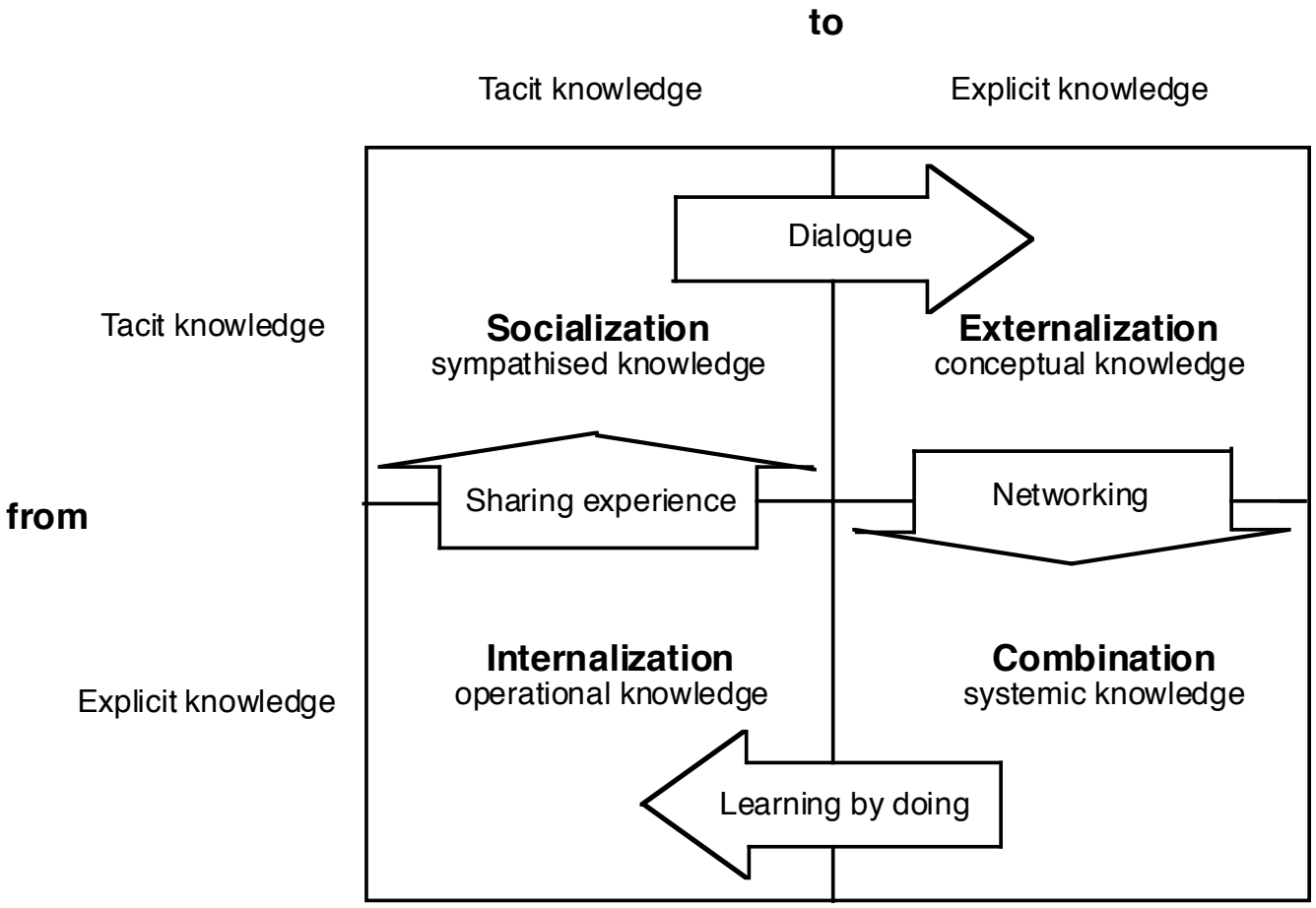

Figure 1: Four modes of knowledge conversion (after Nonaka & Takeuchi, 1995).

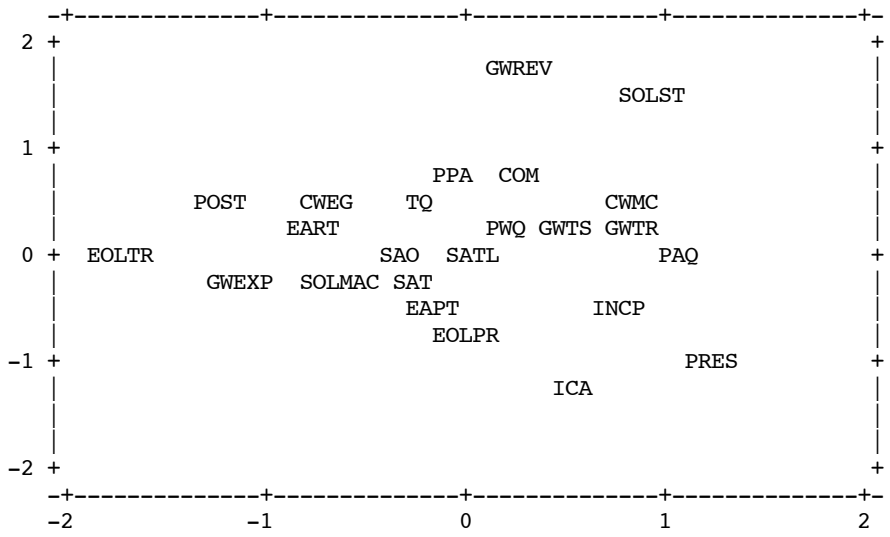

Figure 2: multidimensional scaling of teacher action plan data

1.5	5
1.4	
1.3	
1.2	0
1.1	5
1.0	
0.9	
0.8	
0.7	
0.6	6
0.5	
0.4	0 6
0.3	4 6 8 8 8
0.2	0 3 4 5 6 6 7 9
0.1	5 9
0.0	
-0.1	3
-0.2	
-0.3	4
-0.4	5 9

Figure 3: overall standardised effect sizes

	N I S		L P		D
1.5		1.5		1.5	5
1.4		1.4		1.4	
1.3		1.3		1.3	
1.2		1.2		1.2	0
1.1	5	1.1		1.1	
1.0		1.0		1.0	
0.9		0.9		0.9	
0.8		0.8		0.8	
0.7		0.7		0.7	
0.6	6	0.6		0.6	
0.5		0.5		0.5	
0.4		0.4	0 6	0.4	
0.3		0.3	4 8 8	0.3	6 8
0.2	6 9	0.2	3 4 6	0.2	0 7 5
0.1	9	0.1		0.1	5
0.0		0.0		0.0	
-0.1		-0.1	3	-0.1	
-0.2		-0.2		-0.2	
-0.3		-0.3	4	-0.3	
-0.4		-0.4	9	-0.4	5

Figure 4: standardised effect sizes by control type