

Occasional Paper Series  
Paper 45

*A changing climate for educational research?  
The role of research capability building*

Stephen Gorard

Cardiff University School of Social Sciences

ISBN 1 872330 72X

**Research Team**

Executive Group:

Professor Stephen Gorard (Director)  
Professor Gareth Rees  
Professor John Furlong  
Dr Laurence Moore  
Professor Ken Prandy  
Dr Ray Crozier

Research Staff:

Dr Chris Taylor  
Patrick White  
Helen Taylor (Project Administrator)

**Contact Details**

ESRC TLRP Research Capacity Building Network  
Cardiff University School of Social Sciences  
Glamorgan Building  
King Edward VII Avenue  
Cardiff  
CF10 3WT

Tel. 029 2087 5345  
Fax. 029 2087 4678  
Email. TaylorH1@Cardiff.ac.uk

[www.cardiff.ac.uk/socsi/capacity](http://www.cardiff.ac.uk/socsi/capacity)

# **A changing climate for educational research? The role of research capability-building**

Stephen Gorard  
Cardiff University School of Social Sciences  
King Edward VII Avenue  
Cardiff, CF10 3WT  
02920-875113  
email: gorard@cardiff.ac.uk

## **Introduction**

As part of the Teaching and Learning Research Programme, the ESRC have funded a totally new kind of project, which is likely to be watched with interest by others in social science more generally. This Research Capacity-Building (RCB) project (grant number L139251106) is an innovative attempt to invigorate an entire research field. Among its aims are to support and encourage: the management of complex projects, a widening of methodological approaches, the further combination of different approaches from different contributory disciplines, the melding of theory-building and method, and the creation of new models for transforming findings into usable forms. These aims are not unique to educational research, much less teaching and learning research. They would be seen by many, not least the ESRC (Marshall 2001), as appropriate for other social scientific endeavours such as sociology, psychology, economics and geography. Education is being used as a testing ground for a new approach to capacity building, and whatever the results are they will be a valuable guide for future projects with similar aims in any substantive area.

As the project has only just started this paper can do little more than describe what we intend to do, and perhaps why. The paper begins by rehearsing the background to the call for RCB in terms of a perceived crisis in the quality and relevance of UK educational research (What is capacity-building intended to solve?, p.2). It continues by outlining some responses to the current perceived deficits in educational research - such as greater political control, evidence-based approaches, greater use of statistical techniques, and randomised controlled trials - looking at the likely advantages and limitations of each (Other ways forward?, p.4). We then briefly consider what RCB is (What is RCB?, p.27), and describe how we will approach the task (What we propose?, p.29). The final section considers some of the wider implications of our brief, and some of the difficulties we may encounter (What problems do we face, p.35).

The paper begins by asking for your help in providing practical suggestions before we start our project, and your co-operation once we have started. It needs to be emphasised that our role is to *support* the capacity-building of ourselves and others, chiefly by sharing existing expertise within educational research and across subject discipline boundaries. Our role is *not* to provide the motive force for that improvement, nor to peddle a particular method, or privilege a specific approach. However, we all learn lifelong - about methodological approaches as much as anything else. Knowing more about a wider range of methodological approaches allows greater flexibility in addressing future research questions, but it also allows us to read, understand and where necessary critique, the work of others. We cannot allow

the situation to continue in which different parts of the literature are only intelligible to a subset of researchers. This involves changes of two types – by researchers, the better transformation of results into readable and usable formats, and by consumers of research, a willingness to find out about a wider range of methodological approaches.

### **What is capacity-building intended to solve?**

The origins of the ESRC Teaching and Learning Programme (TLRP) lie, to some extent, in recent critiques of the value and effectiveness of educational research as a contribution to the improvement of teaching and learning in the UK (Hargreaves 1997). It has been argued that 'despite the expenditure of over £65 million of public funding on educational research each year, there are surprisingly few studies which individually, or collectively, contribute systematically to the development of a comprehensive body of high quality evidence' (Millett 1997, p.2). In fact, the value and effectiveness of research as a contribution to the improvement of education has been increasingly called into question (e.g. Hillage et al. 1998. Tooley and Darby 1998). Educational research has been accused of being both 'second rate' and irrelevant to the needs and interests of practitioners. Chris Woodhead, then Her Majesty's Chief Inspector for Schools, claimed to have given up reading research as 'life is too short. There is too much to do in the real world with real teachers in real schools to worry about methodological quarrels or to waste time decoding unintelligible, jargon-ridden prose to reach (if one is lucky) a conclusion that is often so transparently partisan as to be worthless' (Woodhead 1998, p.51). Whatever the actual merits of such points, this 'epistemological crisis of confidence' (Furlong 1996) or 'crisis of legitimation' (Pirrie 2001) is not confined to the UK (NRC 1999, Resnick 2000) nor the field of education. Indeed it is currently characteristic of the relationship between the majority of professions and higher education (and there have been similar changes in the conduct of research in many public services, Dean 2000). However, it is important to note that while a considerable amount has been written about what educational research should now be doing, the actual empirical base on which the original criticisms rests is actually very slight.

At heart, these criticisms address two main issues. The first is the claimed lack of real-world relevance of much research. Much educational research in the UK is not directly transferable into improved pedagogic practice or policy-making, although whether it should be so is a matter for debate. Many of those criticising the relevance of research have a very narrow, usually initial-school, view of what constitutes learning, for example. The second issue is an apparent system-wide gap in expertise in large-scale studies, especially field trials derived from laboratory experimental designs. Over the last twenty years, there has undoubtedly been a move towards much greater use of 'qualitative' approaches (Hayes 1992), even in traditionally numerate areas of educational research (Ellmore and Woehilke 1998). In addition, acceptance rates for 'qualitative' publications are higher than for 'quantitative' pieces, by a ratio of around two to one in one US journal (Taylor 2001). There is a danger therefore of applying different standards of rigour to papers depending on their method, and therefore, presumably, on their referees. However, quantitative work has not stood still, and in the same period techniques for multivariate analysis, of non-parametric data especially, have become considerably more sophisticated. While welcome, these twin developments may have led to a kind of schism because individual researchers tend to specialise in one approach or the other. It is not unusual for one researcher never to have

conducted any form of textual analysis, and for another to admit to not having the least idea what 'multi-level modelling' is about, for example.

Re-reading the findings of the BERA survey of educational researchers from 1978 makes it clear how much more professional the field has since become. Very few researchers then had a permanent job, and even fewer had any post-graduate or professional training in research methods. Most were either fresh graduates or ex-teachers, and several of these willing and able individuals have become the educational research leaders of today. But there are some indications that these researchers - who may now run departments, edit journals and referee grant-proposals - have largely retained the same set of research skills as they started with. There are few indications that these established figures vary their methodological approach to suit the problem at hand rather than seeking out problems that suit their existing methodological repertoire. It is probably the case that we could list many senior BERA members and predict very precisely the methodological approach that *will* be used in their next study. These same individuals are PhD supervisors, and research trainers and models for new generations since they have, by definition, been successful.

The professional situation for researchers is better today - both in terms of tenure and research methods training (at least nominally). However, these improvements have come at some cost. Public expenditure, both directly via grant funding and RAE and indirectly via salaries, on educational research has increased in real terms. This increase has not, in general, been invested in large-scale long-term projects or expensive research. Rather, it has been spread thinly between a growing number of institutions and individuals, whose research culture stems from a previous era in which small-scale work was made worthwhile because small-scale was all that was possible perhaps (at one stage the SSRC, as it then was, was threatened with complete closure). The generation of researchers represented in the BERA study are now, in many cases, the leading 'lights' of educational research today mostly using what is now the standard approach - qualitative analysis with a grand-theory backdrop. This approach is cheap, plausible, attractive to anyone who is not prepared to work with numbers and, ironically, is still seen by many of that same generation as somewhat radical. It is taught to and by successive generations of researchers, who may come to view 'quantitative' approaches done on a shoestring, as consisting solely of surveys, attitude scales, factor analyses, and perhaps the work of the school effectiveness movement. Education as a social science has never found field experiments easy or even apparently useful, while official datasets have only recently become readily available (through a combination of information freedom and information technology), and techniques for their re-analysis are still largely absent.

According to a more recent study of educational research capacity in the UK, there is still a lack of developed research expertise among the people involved, perhaps especially in comparison to other disciplines (McIntyre and McIntyre 2000). The outcome is that a 'large proportion' of the research done (in teaching and learning in this instance) is of poor quality. The skills that do exist are in specific regions and institutions, sometimes represented by one individual (Furlong 2001). These skills tend to be more common in 'qualitative' approaches, which were more popular in the 1990s. Many areas of educational research are now 'dominated by reports of small-scale local studies' (Dyson and Robson 1999, p.vi).

Clearly, to overcome these perceived problems does not mean that we necessarily need more complex techniques. Our approach could start from a consideration of the importance of 'truth' (Bridges 1999),

and a return to a political arithmetic tradition (Mortimore 2000). Clearly also, to repeat and heed these criticisms does not involve an unquestioning acceptance to their validity. We simply acknowledge here the concerns of an important section of UK educational research - including several funding bodies, many practitioner-users, and not a few academics – and treat them as worthy of examination. Our view is that rather than fearing this methodological debate, and perhaps demonising those who seek to take the field forward even when they appear misguided (Ball 2001), we should embrace it as a sign of progress. Lakatos (1998) suggests that it is precisely this kind of methodological debate that characterises an immature research field, and distinguishes it from a mature one that has already passed through the stage. So, an optimistic view would be that, via this recent crisis, educational research may be coming of age.

### **Other ways forward?**

A variety of solutions have been proposed to the problems outlined above, although clearly there is no unanimity that these problems do, in fact, exist. These include increased political control of research and its funding (to ensure relevance), greater use of research syntheses and meta-analyses (to transform findings into policy and practice), simply more quantitative studies (to return the field to a more balanced approach), and greater use of experimental approaches (to provide more rigorous and believable results). Each of these suggestions has at least some merit, but each also involves a few dangers, and has serious limitations if presented as a sole panacea.

#### *Political control and relevance*

We need to consider dispassionately the possibilities and limitations presented by the new climate of evidence-based policy and practice, and the appeal for more experimental-type designs in UK educational research. What we must not do is conflate the issues of relevance, quality and funding unnecessarily. There is a danger that the perceived lack of rigour in current research designs is being used, perhaps unwittingly, to argue for greater political control of research funding, thereby ensuring relevance. However, political control, relevance to practitioners, and even lavish funding do not necessarily lead to high quality research and secure findings (see Gorard 2001a). The situation in UK educational research is serious enough, and the threats of irrelevance or political control of findings real enough, for the research community to take the job of increasing their own research-capacity seriously. Some would say that this may be the last chance for researchers to be allowed to 'police' themselves.

Allowing political bodies, and not academic researchers, to decide on a research agenda may produce greater short-term relevance but is almost certainly not going to produce research of higher quality. There may, in fact, be a tension between looking for rigour and looking for relevance in research. Central planning has failed to produce better research in other fields (Hammersley 1997). Many researchers who have worked on consultancies, contract research and evaluation studies will have experienced the pressure, subtle and not so subtle, put upon them to produce results in accord with some pre-determined plan. It is as though 'research' is being conducted to find evidence for already existing agenda. Perhaps when politicians with legal experience talk about evidence-based policy they do not mean making policy based on research evidence about what is likely to be most effective or the fairest (the social science definition). An alternative, and equally plausible, interpretation (the legal definition) would be that evidence was what a policy-maker sought to help establish a case for a policy.

It is therefore important for all concerned to decide which of the two versions is being supported in the push for evidence-bases. They are incompatible, and while the confusion remains there is a clear danger that research is being subverted to a role of legitimating policy (Levacic and Glatter 2001).

Partly as a response to the UK debate about the value and relevance of academic educational research, government appointed bodies have also tried to move some research funding to practitioner groups. The results have so far not been very impressive (although it should be stressed that the examples used below come from the early stages of the scheme). The Teacher Training Agency (TTA), for example, award small grants to teachers to carry out research 'relevant' to their needs. Judged on the basis of the best of the pilot studies, the research so funded is often actually not research at all, not being aimed at producing knowledge based on the collection of evidence. It often simply describes current practice or uncontrolled attempts to change it. Not all studies produce a report, but even the rest omit crucial details such as what they had done, or what their evidence was. It has been claimed that their apparent conclusions are mainly repetitions of previously held opinions (there being no sign of the surprise that is the hallmark of real discovery), but their 'lack of critical scrutiny also allows the presentation of questionable ideological views about the nature of practice under the banner of scientific research' (Foster 1999, p.396).

Perhaps the best such study concerns an 'experiment' on the teaching of mental arithmetic (TTA 2000). In the school concerned the 1998 Key Stage One cohort (54 pupils) were taught mental arithmetic using a new method. A higher proportion gained level 3 than in the 1997 cohort (55 pupils) taught using another method. Thus, the researchers conclude that the new method is better, and the TTA have supported this view by publishing it. Despite being the best study of the first 26 funded (by presenting evidence and method for example, and having a reasonable number of cases), this is not good research. The findings are simply not safe. The cohorts were different and unmatched, the proportions gaining level 3 are growing year by year nationally anyway, the two groups sat different test papers, the teachers were not matched for the different cohorts (the Maths co-ordinator taking the second group), and there is no consideration of either the Hawthorne or experimenter effects (see Gorard 2001b).

As another example, the DfEE apparently spent nearly £4 million on the Hay/McBer research into teacher effectiveness. The result was a description, and by implication a prescription, for what effective teachers actually do when teaching (Hay/McBer 2000). The relevance of this work cannot be doubted. What is more doubtful is the claim to knowledge stemming from this colossal (in educational research terms) expenditure. And an important point to note is that the work has not peer-reviewed in the usual sense, like the 'Models of Headship' work also funded by the DfEE (as it then was) but which has not been made publicly available.

In essence, the researchers are saying that they have described, via a range of data collection techniques, what 96 effective teachers are like. However, even assuming that the researchers can successfully identify an effective teacher this description is of little practical use. If these 96 teachers tend to dress smartly for example, then an argument could be advanced that dressing smartly can lead previously less successful teachers to improve. This is, of course, nonsense but it is very popular kind of nonsense (Davis 2001). The effective teachers were largely selected by asking other people who is 'effective', and so the ensuing description of the attributes that led to their identification is that of other people. The study tells us what an unpublicised group of 'other people' think makes a successful

teacher. It is no surprise therefore that, like prescriptions for school improvement, the ensuing model contains elements of tautology, including the obvious and the nebulous.

In fact Hay/McBer defined effectiveness both in terms of ratings and observations, and from pupil progress scores. In order to arrive at the 96 cases, it was only possible to use schools with high quality pupil records and other datasets. The population for the study is therefore not, as it appears at first sight, all schools but all schools with high quality datasets (i.e. all those with a non-zero chance of being in the sample). It is perfectly possible that such schools, and the teachers within them, will tend to be different from the rest in important ways.

Even given this restriction on data quality, the number and range of actual datasets encountered in the sample schools were considerable (Reynolds 2000). Half of the schools had unique home-made schemes for calculating value-added scores, and the remainder used various scoring systems such as Yelsis or the optional QCA SATs. Most also used some socio-economic background or contextual variables, but again these were not consistent from school to school. Making the best use of the compromise data available, Reynolds (2000) converted the pupil gain score for each teacher into categories such as A+ (very effective) to C (poor), and the judgement of these categories was based on pupil scores for each teacher relative to context and to prior attainment.

The Hay/McBer study also grouped these teachers, but not using the same categories preferring instead a range of 'Outstanding+' to 'Poor'. These judgements were made on the basis of observations and the rating of others. Teachers judged ineffective or poor were dropped from the original sample of 172, leaving only 128. So now the sample, and its population, consisted of teachers from schools with adequate datasets, not rated ineffective by others. The ratings score was correlated with the value-added score, yielding a coefficient of +0.43. This means that if, for the present, we accept the validity of both approaches then only 18% of the variance in these already heavily edited groups is common to both. If outstanding teachers are meant to lead to greater progress for their pupils then this 18% is prime facie evidence that one or both of the classification systems is heavily in error. Given that the 'Reynolds' classification takes into account contextual figures whereas the 'Hay/McBer' does not this may suggest that it is the latter which is more at fault. However, rather than conclude anything like this the study edits the 'sample' again. It only proceeds to analyse and characterise those individuals whose ratings intersect to a large extent. Put another way, 32 teachers who have A+ pupil progress scores but were not considered to be good teachers by others, or who had weak pupil progress scores but were considered 'outstanding' by others, were dropped (and remember that purportedly poor teachers have already been dropped). Not surprisingly the correlation between the two classifications for the remaining 96 cases rose to +0.79, but it is inappropriate for the study to even attempt to cite this figure. Running a correlation, removing cases from the awkward diagonal, running it again and then reporting the final correlation coefficient is not good science.

Returning to the already heavily-edited sample of 128, for which Reynolds (2000) reports a 'moderate correlation' between the two assessments of teachers effectiveness, 51 of these were rated outstanding by Hay/McBer, while 43 were given very effective grades by Reynolds. Only 28 of these individuals were in the intersection. Collapsing the various categories in the two different scales into two each - whether outstanding or not - leads to Table 1. Therefore, only 55% (or 28/51) of teachers rated outstanding by Hay had better than typical pupil progression. This is not sufficient to enable the study to

claim that they can describe, and perhaps prescribe, the characteristics of an effective teacher. Trimmed to this summary, the report is basically a list of the attributes people already think a good teacher has.

Table 1 - Crosstabulation of teachers effectiveness: Value-added versus Hay/McBer

	Value-added B/C	Value-added A	Total
Hay Outstanding	23	28	51
Hay Typical	62	15	77
Total	85	43	128

While it might be as unfair to single out the studies above for their problems, as some people consider it was to single out academic studies in the Tooley report, there is no sense here that increased political control and relevance alone are likely to lead to higher quality research. Perhaps what researchers actually need is more independence from political pressure. They need to be able to reach conclusions that are as secure as possible with regard to the evidence, but not the consequences in particular situations. Evidence gives us a basis for belief and so for decision, but the political decision remains a largely separate stage, for depending on the starting assumptions the same evidence could be used to justify different decisions - as was made clear when an internet search engine threw up two adjacent references to some recent work on school compositional effects. One website was for the Campaign for State Education, who used the findings to defend the performance of comprehensive schools. The other was termed the 'Reagan page' and used the same evidence to argue for more diversity among schools in the US. Both uses were coherent, for the evidence was largely neutral, and the difference lay chiefly in the starting assumptions of the political commentators.

The government wishes political pressure on researchers to increase, to control research both financially and intellectually, but perhaps the greatest pressure is coming from researchers themselves. We seem to forget that no scientist is morally obliged to come to an unwarranted conclusion or inference. There are whole fields of endeavour in educational research in which the 'researchers' take sides before any evidence is collected and where the evidence collected is knowingly affected by this bias (Gorard 2000a). Of course this problem is not unique to education. [A professor of sociology recently came to me with a paper fully written, presenting the damage done to local communities by a particular policy decision, but with no data. He asked me if I had any figures, from the Labour Force Survey for example, which would back up this point. He is not, in my experience, unusual in this approach of 'premature adjudication', and he is successful, mentoring new staff, supervising PhDs, and being awarded large amounts from public funds to pursue this 'research'].

Surely, unless researchers can greet all possible outcomes of their empirical investigations with at least equal respect, and they are prepared to be surprised by what they find, and they apply the same standards of rigour to other people's work regardless of whether they approve of the findings or not, then they are not behaving as researchers but as political agents. When we as peers say 'I really like this work' we are all too often saying that we approve of the findings rather than admiring the rigour of the analysis or the replicability of the data collection or the logic involved in drawing the conclusions. Some 'researchers' explicitly advocate taking sides *before* collecting the necessary empirical evidence (Griffiths 1998). But the researcher cannot afford to 'take sides' with anything but the truth (Bailey 2001). When a researcher has an agenda, such as anti-racism, this should make no difference to the research. The key problem with 'racism' as a phenomenon is that it concerns an important issue (which

is why it is being researched) and that it is unjustified (by the truth). Since racism is unjustified by definition, the researcher therefore does not need anything other than a desire to find and publish the truth, which is presumably what other researchers do anyway. Similar, but more complex, arguments can be advanced to show how other apparently radical stances, such as 'feminist' researcher, are also entailed and therefore essentially unnecessary. Research cannot be simply subordinated to the needs of policy-makers and political agitators without being in danger of becoming contract 'research' which may be used to justify already-prepared programmes of action (to benefit either the government or a political pressure group).

While this account of political control of research is negative in its assessment, this should not be seen as an argument for total freedom. As with most things the need is for balance. Research fields with no political interference tend to become 'chauvinistic', in the sense of resisting alternatives to the status quo (Feyerabend 1993). Perhaps that is what has happened in UK educational research, and perhaps the current threat to their freedom will encourage academic researchers to embrace with caution the other improvements to theory, method and dissemination suggested here. The British educational research community certainly needs to start policing itself better before someone else is appointed to do the job for them (Brown 1998).

#### *Evidence-based policy and practice*

Of the wide variety of responses which have been proposed to the problems above (Furlong 2000) the most prevalent is probably the call for research which focuses explicitly on 'what works'. Such research could then be used as a basis for establishing 'evidence-based' practice where pedagogical and other decisions are guided by nationally agreed 'protocols' (as in the field of medicine, Department of Health 1997). Syntheses of high quality studies are used to produce the findings, which are then 'engineered' into practice. The assumption is therefore not that good evidence has not been provided by previous work, but that it is difficult to see its pattern without systematic evaluation (the work is fragmented, Pring 2000), and impossible for it to have an impact on policy and practice with re-engineering. Simply publishing results is not enough. The beauty of this solution is that it apparently addresses issues of both relevance and quality. It can therefore be justified on solid practical grounds. For example, in a review of administering albumin to humans, Roberts (2000) concludes that it 'provides a strong argument for preparing scientifically defensible syntheses of the evidence from randomised controlled trials in medicine, as well as in other important areas of social policy, such as education' (p.235). The significance of this is that if albumin administration had ceased in the UK when doubts were first raised, this synthesis suggests that around 10,000 patients who died may have been saved. Relying on theory rather than heeding the warnings from trials led to needless loss of life.

Measuring output effects is the 'bottom-line' in business, as well as medicine, and perhaps crime prevention studies. The emphasis is on whether an intervention works, not why (Brighton 2000). This approach, at least implicitly, sees large-scale randomised controlled trials as the ideal form of evidence, which a systematic review further improves by minimising bias through selection and omission, leading to safe and reliable results (Badger et al. 2000). The call has also been made for research to be concentrated in fewer centres to encourage high levels of skill, the establishment of good networks of users (for interactive social science), and so prevent fragmentation and the poor quality of some existing research (Pring 2000). While these approaches are promising, and will be viewed with interest, they are unlikely to provide complete solutions. For a start, someone is still required to undertake the primary

research that is being synthesised. Even in medicine, which receives a lot more funding than educational research, this approach is being criticised (Hammersley 1997).

Also, while plausible, this approach does face technical difficulties that are not always highlighted by its advocates. Replication of studies, of the kind needed for syntheses and meta-analyses, are not currently encouraged by the funding or publishing mechanisms in the UK. This means that results are usually equivocal in practice, especially as educational research in particular is hard to simplify and do justice to. 'Qualitative' evidence is largely ignored by this approach, which is particularly wasteful (Levacic and Glatter 2001). Systematic reviews can therefore be misleading by hiding details, and privileging trials even where considerable evidence of other forms contradicts them. This has led to false conclusions that are just as important, in reverse, as those claimed for the evidence-based approach. For example, by conflating different periods and schedules for a synthesis, it has been concluded that a short treatment for alcohol abuse is as effective as a long one, and that sexual abstinence lessons for 13 year-old boys actually lead to more sex. Both of these results have now been attributed to bias in the meta-analytic process (Speller et al. 1997). Meta-analysis is usually conducted on the assumption that all studies involve samples taken from the same population and with a constant effect size (Field and Wilkinson 2001). Since these are, in fact, unlikely to be true the misapplication of meta-analysis increases the chances of incorrectly detecting an effect, especially where the number of studies is relatively small.

Clearly no-one is going to present a coherent argument *against* the use of evidence in forming policy or practice. But evidence-based approaches imply more than that. Their chief problem is that education does not have a single agreed indicator of what works. In medicine, survival or clear-up rates are relatively unambiguous. In criminology, crime reduction plays a similar role. Perhaps education is more like health promotion than either of these. Health promotion is having difficulty showing any effectiveness in the existing Cochrane model. This may be because there is little effect, or it may show that educational impacts require different kinds of measurements. Even in medicine, it is already clear that the quality of the research question and of the intervention must be as good as that of the trial methodology. The development of these interventions and the measures of them must be based on theoretical models (see below) assisted by 'qualitative' studies to suggest why the intervention might work. All trials (even of the 'black box' type) are as specific in time and place as their intervention. Thus, to learn from a trial we need an explanation of its effect, and cannot rely solely on what works. As in the Hitch-hikers Guide to the Galaxy, knowing that the answer is '42' merely leads us to further questions.

Evidence-based approaches are an important way forward for educational research, when taken in combination with other approaches. In some cases, however, the political criticism of existing work is based on a misunderstanding of the role of research, or even a dislike of the results (or the method). A real evidence-based approach to policy-making would require also a fundamental change in the way politicians make and carry out policy that is unlikely to happen in fact (Pirrie 2001).

### *Simply more statistics?*

If political control or evidence-bases are not sufficient to overcome the apparent crisis in educational research, perhaps the solution is simply a technical one of increasing the role of large-scale studies and techniques for the analysis of complex datasets. Some people may have suggested that the reason we require more statistical ('quantitative') studies is that this form of evidence is intrinsically preferable and of higher quality than other forms. We feel that this is completely the wrong way of looking at it. On the

contrary, one reason to encourage a greater awareness of statistical techniques among researchers is that quantitative work is currently often very poor. Our motive is not chiefly to increase the amount of such work but to raise its level. Another reason is based on our desire to see greater use of combined approaches (see below) which, of necessity, requires the majority of researchers to have a working knowledge of basic numeric techniques.

Our belief therefore is that all researchers should use numbers routinely in their research (even if only as 'consumers' of the quantitative research of others). The first and most obvious point here is that the process of research involves some consideration of previous work in the same field. All researchers read and use the research of others. So they need to develop what Brown and Dowling (1998) refer to as a 'mode of interrogation' for reading and using research results. If they do not have any understanding of research techniques involving numbers then they must either accept all such results without question, a very dangerous decision, or ignore all such results, a very foolish decision. In practice, many commentators attempt to create a middle-way of accepting some results and rejecting others even though they do not understand how the results were derived. This usually means that results are accepted on the basis of ideology, of whether they agree with what the commentator wants. This is both dangerous *and* foolish.

Another reason why all researchers are likely to need numbers, irrespective of their primary method, is that many of the large datasets available as context information for any study are numeric. The use of secondary data to help create or identify an appropriate sample (perhaps via stratification), to describe the pattern or problem to be explored by other methods, or even as a method in its own right, is growing (Gorard 2001b). This is a trend encouraged by the funding councils, and welcomed by us. It allows cumulation, and helps prevent the waste of resources involved in attempting research to explain non-existent patterns or problems. Accounts based on numeric *and* 'qualitative' analyses too often lead to an apparent contradiction (in Riddell 1992, p. 46) because, too often, the qualitative analyses are attempting to explain a situation based on flawed numeric reasoning. Existing statistics, whatever their limitations, provide a context for any new study which is as important as the 'literature review' and the 'theoretical background'.

The range and scale of existing datasets is phenomenal, and growing all of the time. Few will ever be exhausted as research resources, and the creative combination of data from two or more can lead to considerable originality. If the questions we want answering are already addressed by an existing dataset (the Individual Student Records, the annual school census, the population census, the Labour Force Survey, Household Panel Survey, National Child Development Study etc.), then we are likely to obtain our answers quicker, cheaper and better from them than by conducting our own research. However, we generally have no agreed methods for dealing with these large, and often complex, secondary datasets (Gorard and Taylor 2002a). There are currently debates over the way to measure trends over time, differences between places, and how to deal with hierarchical data for example. These debates need to be pursued with vigour so that relatively standard protocols can be produced for general researchers to use when simply wishing to conduct a 'smash and grab' on existing data in preparation for a new study.

A problem we face, perhaps most common in the UK sociological tradition, is that some researchers simply reject all numeric evidence and its use (what Mortimore and Sammons, 1997, call 'crude anti-quantitative attitudes', p.185). Having realised that numbers can be used erroneously, sometimes even

unscrupulously, these researchers simply reject all numeric evidence. But, as Clegg (1992) points out, we know that people sometime lie to us but we do not therefore reject all future conversation. The danger for 'qualitative' research conducted in isolation from numeric approaches is that it can be used simply as a rhetorical basis for retaining an existing prejudice. Without a combination of approaches we are left with no clear way of deciding between competing conclusions. Our argument is therefore that numeric evidence forms the basis of good qualitative studies, and can be used to test its findings (and vice versa of course - the middle-way, see Gorard 1998). So much is this so that it is not clear that the distinction between the two forms of evidence is a useful one.

The above comments should not be seen as a ringing endorsement of existing 'quantitative' approaches, nor as a plea for more of the same. Another problem, perhaps more common in US psychology, is that there has been a tradition that only numeric data is of relevance. There is sometimes a tendency to count or measure everything, even where this is not necessarily appropriate (as with some attitude scales for example), and one outcome is that statistical analysis may be done badly. Forms of evidence not based on numbers are nevertheless despised, while evidence based on numbers tends to be accepted less critically. Part of the problem here may be the 'cronyism' among reviewers that in-depth knowledge of advanced statistical procedures tends to generate, which leads to poorly explained and over-technical reports (where incomprehensible software-generated variable names are used routinely in descriptions of multi-level modelling, for example).

The increasing quality and availability of computer software packages for statistical analysis allows more and more complex statistical models to be built and used, so that in the end most consumers of educational research simply cannot, or would not wish to, comprehend them. Even those who work on such high level models have trouble transforming their findings into a package that does their analysis justice but also makes any sense to practitioners and policy-makers (see Goldstein et al. 2000). This means that the 'average' consumer of research has to either implicitly accept the findings, or reject them as incomprehensible. Linked to the greater use of computers is the shotgun or dredging approach to analysis in which multiple exploratory analyses are run with the same set of data. As well as liberating us from the drudgery of multiple calculations the computer has therefore increased the frequency of the 'blind or mindless application of methods without regard to their suitability for the solution of the problem at hand, or even in the complete absence of a clearly formulated problem' (Pedhazur 1982, p. 3).

Statistical procedures are ideals, but several studies of actual behaviour have observed different common practices among researchers. 'Producing a statistic is a social enterprise' (Gephart 1988 p.15), and the stages of selecting variables, making observations, and coding the results, take place in everyday settings where practical influences arise. The divergence between ideal and actual is probably increasing because of the increased accessibility to statistical software packages, and a tendency to see these as 'expert systems' rather than convenient calculators. Statistical packages are making decisions for us that we may not even be aware of (through default settings). The possible dangers of this are increased because for some, statistics have an under-stated rhetoric of their own, able to persuade specific audiences of their objectivity (Firestone 1987), which perhaps helps to explain why so few academic disputes over figures, and subsequent corrections by authors, appear in UK educational literature. The average researcher may be easily fooled by large numbers, confused by probabilities,

prone to the fallacy of *post hoc ergo propter hoc*, and, without expertise of their own, led (and perhaps misled) by authorities (Brighton 2000).

Where statistical techniques are involved in research there is little general understanding of their strengths and limitations (Field and Wilkinson 2001). Analysis of variance, designed for agricultural applications, is based on an assumption that there is no measurement error involved, whereas it is used routinely in education and psychology where the measurement error is considerable. The least-squares model of regression in common use faces technical deficiencies, and has had limited success. It has never predicted anything like an eclipse, or even a thunderstorm (unlike fluid dynamics, for example, which takes existing known interactions onto account, Brighton 2000). In addition to incomprehensibility, the move towards more and more complex forms of multivariate analysis is increasing the schism noted at the start of this paper, and leaves 'bread-and-butter' issues such how to measure differences between things unresolved (where only primary school arithmetic, but advanced level logic, is needed). There is also a danger that the same people who might argue that a correlation is not the same as causation, have paradoxically seen complex models based on correlation as being causal models (see below). This is a real danger of complexity, along with the reviewing problems that it entails. Hierarchical linear modelling, or multi-level modelling, is clever and complex, but also difficult to explain, incompletely theorised (what determines the levels and their relationships, can variance be partitioned, or does Hebb's rectangle apply?), and loses the concept of an individual predicted score that simpler methods allow. Above all, how much practical benefit has been generated thereby? Put another way - what secure knowledge has it generated that would not have been possible anyway? The focus of school effectiveness studies (the main use of HLM/MLM so far) has been very technical. Its findings have not always been transformed into anything usable (recall that a considerable amount of the value-added work actually going on in schools is *not* based on these techniques). It can be seen by teachers as being imposed on them from outside, and little is actually done with it in schools once the researchers have left (Saunders 2000). Perhaps 'we don't need more complex analytic techniques, we need better data collection' (Brighton 2000, p.135).

There are a variety of other problems facing increased use of statistical procedures in educational research. These include the problem of error propagation (and even ill-conditioning) in calculations with the level of representational and measurement commonly found in all social sciences (Gorard 2001b), the problem that the difference between real-world logic and mathematics can lead to confusion (see below), and the debate about the routine flouting of assumptions concerning levels of measurement or distributional parameters. The single biggest problem might be the continued over-use of null hypothesis significance testing (NHSTs). These tests were designed to separate the likely influence of chance in selecting a random sample from all other influences on our results. They generate the probability (p-value) of a result as extreme or more extreme as the one obtained, if the only reason for the result is the variation introduced by the random sample. Where the sample is not random or where a population is used then this probability is meaningless, and NHSTs are therefore useless (for that is *all* that they do). Even where a random sample is used, the variation due to sampling is likely to be so small in comparison to that due to design bias, measurement error, dropout, non-response and so on, that NHSTs are still virtually useless. Although there are suggestions to replace these p-values with standard errors, many of the same problems would continue to apply. It is not clear why we should use standard errors anyway. They are not used in business reports, or examination grades, for example, where they

might be just as appropriate. In real-life the best estimate is our current score for any measurement (while we should treat all such scores with caution).

Therefore, traditional statistics will continue to be useful in modelling the process of education, but like political control and evidence-based practice must be treated with caution, and seems unlikely to provide the answer to all of our problems. Perhaps we should turn to the process of data collection, and the design of new forms of experiments.

### **Do we need more experiments?**

In many ways the experiment is seen as the 'flagship' or gold standard of research designs. The basic advantage of this approach over any other is its more convincing claim to be testing for cause and effect, via the manipulation of otherwise identical groups, rather than simply observing an unspecified relationship between two variables. In addition, some experiments will allow the size of any effect to be measured. It has been argued that only experiments are thus able to produce secure and uncontested knowledge about the truth of propositions. Their design is flexible, allowing for any number of different groups and variables, and the outcome measures taken can be of any kind although they are normally converted to a coded numeric form. The design is actually so powerful that it requires smaller numbers of participants as a minimum than would be normal in a survey for example. Educational research has, for too long, relied on fancy statistical manipulation of poor datasets, rather than such well designed studies (FitzGibbon 1996).

The experimental method can also be extremely useful to all researchers even if they do not carry out an experiment. Knowing the format and power of experiments gives us a yardstick against which to measure what we do instead, and even helps us to design what we do better. An obvious example of this is a thought experiment in which we can freely consider how to gain secure and uncontested knowledge about the truth of our propositions without any concern about practical or ethical considerations. This becomes the ideal, and it helps us to recognise the limitations of our actual approach. Another example is a natural experiment where we design an 'experiment' without intervention, using the same design as a standard experiment but making use of a naturally occurring phenomenon (Gorard and Taylor 2002b). Natural experiments are going on around us all of the time, when interventions occur as part of the normal policy process. If one local education authority changes its practice in some way then it can be construed as the experimental group and the remaining authorities as controls in a natural experiment. In fact, much social science research is of this type - retrospectively trying to explain differences between two groups. This is inferior in terms of validity to a true experimental design but much more practical. Knowing how an experiment works is important because it enables us to see how far a natural experiment is from that ideal.

The logic of an experiment relies on the only difference between the groups being due to the treatment (intervention), and in this case the experiment is said to lead to valid results. There are several threats to the validity of experiments. An often cited, but still useful, summary of many of these potential threats comes from Campbell and Stanley (1963) and Cook and Campbell (1979). Also, to consider long-term outcomes is expensive and not attractive to political sponsors (who usually want quick fixes). A danger for all educational research is therefore a focus on short-term changes, making the studies trivial

rather than transformative (Scott and Usher 1999). Given these and other potential limitations of experimental evidence (Adair 1973), this ideal, the supposed flagship of social science research, is often far from realisable. There will always be some room for doubt about the findings, even from a properly conducted experiment. It is important, however, to note two points. First: there are some things we can do with any design to counter possible contamination, hence the importance of properly constructed protocols. Second: the experiment remains the most completely theorised and understood method in social science, and it has led to considerable research cumulation in many fields of endeavour of the kind that other, perhaps weaker, designs have yet to achieve.. With its familiarity comes our increased awareness of its limitations, but other and newer designs will have as many and more problems. Worse, they will have dangers and limitations that we are not even aware of yet.

The biggest problem in using experiments comes from their chief source of strength - the level of control of the research situation possible in a laboratory. Traditionally experiments have been conducted in laboratory conditions, following a supposedly natural science model. A laboratory allows the experimenter to control extraneous conditions closely, and so to claim with more conviction that the only difference between the two groups in an experiment is the presence or absence of the treatment. This level of control often leads to an unrealistic setting and trivial research questions. It has been said that a series of experiments allows us to be more and more certain about less and less (in Bernard 2000). In fact, although it may be desirable in research terms, this level of experimental control is usually absent when confronting research in real-life situations. For example, it is not possible to allocate people to groups randomly to investigate war, marriage, employment or imprisonment. In addition, it is actually the control of the experiment by the researcher that can lead to self-fulfilment (delusion), or selective bias in observation. There can also be ethical problems in deceiving participants since, even where the treatment is non-harmful, it is usually necessary for the participants to be ignorant of the purpose of the experiment.

Therefore the laboratory experiment is akin to an ideal. It is how we might like to conduct research to get clear answers about the implications of actions in education. Even if we never conduct an true experiment, knowing what it would have been is an important yardstick to evaluate what we do. Everything that has been said about the problem of internal validity in experiments applies with even greater force to all other, perhaps less well-theorised, designs. If an experimenter, while trying to be neutral, unwittingly conveys demands to participants in fairly meaningless laboratory tasks, then imagine the likely effect of an interviewer in personal communication with a interviewee for example. If an experimenter unwittingly makes favourable mistakes in noting or adding up a simple data collection form, imagine the level of bias possible in interpreting the findings from a focus group discussion. If we consider the ways in which our actual designs are like a true experiment it allows us a glimpse of their considerable imperfections and keeps us appropriately humble. It also helps with future research synthesis (see above) by giving us a standard against which to compare all studies. 'Thought experiments' are now widely used in science. Thought, or fantasy, experiments are quick, cheap, and have no ethical problems (since we have no intention of carrying them out). We can think the unthinkable by imagining what a true experiment would be like for our area of investigation, and then compare the actual and ideal designs to help show up the defects.

A more common (though currently still far from popular) form of experiment in educational research is the field trial, or naturalistic field experiment. The most obvious way in which field trials differ from

laboratory ones is that they tend to use existing groups as the basis for treatment (Hakim 1992). These 'quasi-experiments' therefore do not use random selection or allocation but recognise natural clusters in the population. It is just about impossible to allocate students to teaching processes (such as schools or classrooms) at random. What is possible is to use existing teaching groups and vary the treatments between them, using statistical procedures to try and iron out the differences in results due to pre-existing group differences. This approach gives the experiment a lower general level of internal validity, but because the setting is more realistic than a laboratory the external validity (relevance to real-life) is probably greater. These designs may exhibit less rigour, with no control group or else using self-selecting clusters from the population (and all too often they test incompletely thought-out concepts and questions, Brighton 2000). They therefore require very clear logic in the evaluation of their results and the consideration of alternative explanations to be convincing.

The biggest challenge facing the increased use of experimental designs in educational research is, however, an ethical and not a technical one. Of course, ethical issues do not only apply to experiments. Ethically, the first responsibility of all research should be to quality and rigour. If it is decided that the best answer to a specific research question is likely to be obtained via an experimental design for example, then this is at least part of the justification in ethical terms for its use. In this case, an experiment may be the *most* ethical approach even where it runs a greater risk of endangering participants than another less appropriate design. Pointless research, on the other hand, remains pointless however 'ethically' it is conducted. Good intentions do not guarantee good outcomes. Such a conclusion may be unpalatable to some, but where the research is potentially worthwhile, and the 'danger' (such as the danger of wasting people's time) is small relative to the worth, the conclusion is logically entailed in the considerations above (see Gorard 2001b).

Good experimental designs testing quite narrowly defined hypotheses (to minimise confounding variables) have considerable power, especially as part of a larger cumulative programme of research via replication, expansion and verification of the findings. Above all they can help us overcome the equivalent of the potted plant theory which is distressingly common in educational research, policy-making and practice. This theory suggests that if good schools have a potted plant in the foyer then putting a potted plant in the foyer of other schools will lead to an improvement in their quality. Unless we intervene or rigorously monitor the effect of natural interventions we can never be clear whether our observations of patterns and apparent relationships are real or superstitions. As long as we remind ourselves that the power of experiments comes not just from their design but also from the significance of the problems they are deployed to solve, then the planned growth in experimental studies in education (and public policy more widely) can be seen as valuable and worthy.

The major claim made for the use of experimental designs (Fisher 1935) is their ability, used correctly, to uncover and test causal mechanisms. No other research design has this ability (Gorard 2001e). Field trials (non-laboratory experiments) are also apparently persuasive to the public and policy-makers, and given the recent development of new techniques and protocols (for example from health promotion studies in schools) they are more applicable to education than ever before. Although causation is little taught in research methods courses, it is often talked about negatively in the sense that association, correlation, plausibility and so on should not be treated as causation (Gorard et al. 1998). This lack of ability to test causation in non-experimental designs is a great drawback for educational research, and

means, almost by definition, that many promising avenues lead nowhere of any practical use. The results they generate can always be gainsaid if expressed as causal models, and can be ignored if not.

Cause:effect models abound in educational research. Other than in purely descriptive work (e.g. 'the achievement gap between boy and girls in 1999 was 17%'), a research report that did not at least imply a causal model might look rather odd. Causes are central to our notion of understanding why things work as they do, but they are just as central to the less sophisticated 'what works' approach (e.g. 'become an effective teacher/school by doing what effective teachers/schools do'). Yet despite this prevalence, social science research methods courses and textbooks tend to overlook the discussion of causal models completely, or else prepare the novice researcher simply with the negative advice that a correlation is not the same as causation. If, over time, the income of the Archbishop of Canterbury tends to rise in line with the street price of cannabis this is not evidence that the Church of England makes money from drug-dealing. If weather forecasts were able to predict thunderstorms with perfect accuracy would that make them the cause (Collins 1985)? In these standard accounts, everyone is reminded therefore what is not a cause, and what a cause is not. In some methods books there is a section on the potential and limitations of experiments which points to their unique selling point - the claim to be a direct test of cause and effect. But this is a scarce and recent phenomenon in social science outside psychology. In general, the concept remains untaught and undiscussed (Salmon 1998).

Some research is, and should be, solely descriptive. It is anyway an essential first step to doing exploratory work, since 'before asking why we must be sure about the fact' (De Vaus 2001, p.2). It is, in my opinion, far too common that researchers set out to explain and explore a phenomenon that does not actually exist (Gorard 1998, as cited in Hillage et al. 1998). Recent examples include attempts to explain: the school-mix effect; the growing gender gap in attainment, and increasing socio-economic segregation in school compositions. The fact that we can create a plausible theory to explain imperfectly understood notions such as these is not evidence that they must exist. Such research should, rather, routinely start from a re-analysis of relevant existing datasets, and base the ensuing exploration on the patterns uncovered in the preliminary work (see above).

In the exploratory phase of a study the role of 'unfettered' theory is limited. Whatever methods of data collection and analysis are used, the subsequent theory is an attempt to reconcile empirical findings with pre-existing 'common ground'. This theory is, either explicitly or no, a causal theory (interpreting the term 'cause' in its widest possible sense), since it suggests how the empirical findings arose. The underlying causal model is perhaps clearest when the research is based on an experimental intervention (although implicit even in the most complex designs). This is not a moot point, since the difference between experimental and other forms of evidence can be crucial (Badger et al. 2000). It can save lives (Roberts 2000). Its advocacy by some commentators rests largely on this notion of causation, which makes it timely, for a variety of reasons, to consider here the nature of causal modelling in rather more detail than we usually do.

The paper proceeds by considering three positions in relation to causal models - that they exist, that they do not exist, and that they exist alongside non-causal phenomena. It shows that there is no logical or empirical reason to reject any of these positions, but suggests that educational researchers, by the nature of their remit, are committed to the first. The paper continues by outlining some of the desirable

characteristics of a 'good' causal model, and their relevance for the future of publicly-funded educational research.

Are outcomes caused?

It is not possible to detect a cause empirically or prove that one exists philosophically. We can never directly sense a cause. We merely induce their existence from our experience of the association of two or more events, and this is nothing more than a habit of mind - immutable though it appears (Hume 1962). A very similar process is observed in both classical and operant conditioning, where the association of two things leads the conditioned subject to behave in the presence of one thing as though it implied the presence of the other. Yet unlike conditioning in dogs or pigeons, the process of induction has been presented as the chief criterion of demarcation between what is considered 'science' and what is not. This is why, despite important developments by Kuhn (1970) and then Lakatos (1978), it has tended to remain the 'skeleton in the cupboard of philosophy' (Russell, in Ayer 1972). Our notion of cause is little more than a superstition. Alternative criteria for the definition of scientific, as opposed to non-scientific, endeavours have been suggested (e.g. by Popper, in Magee 1973). The problem with these is that, despite the claims of their advocates, they do not remove the problem of shaky philosophical foundation of induction.

For example, when observation leads us to question a belief because it brings two beliefs into contradiction we tend to stick with the most familiar of the two concepts (Goodman 1973), which suggests that Popper's notion of falsification does not actually eliminate inductive logic. To use Popper's own example. No number of observations of white swans can prove that all swans are white, for even if all white swans could be accounted there may be other swans that are not white. This, in essence, is the problem of induction and therefore of causation. For, however often two things appear together (whiteness and 'swanness'), they cannot *prove* a link. On the other hand, as Popper observes, only one observation of non-white (black perhaps) swan is needed to falsify the proposition that all swans are white. He suggests therefore that scientists proceed not by trying to prove their propositions, but by falsifying them. This is one basis for the purported difference between science and non-science. The problem with this is that Popper, and his advocates, are ignoring the crucial distinction between formal and real-world (henceforth 'Aristotelian') logic. In formal logic, a contradiction such as 'A entails B' and 'Here is an A which is not B' cannot be explored further. The contradiction shows that there is a flaw somewhere in the prior logical chain, perhaps in one or more of assumptions, since both statements cannot be true. Logic does not help us find the flaw (any more than mathematics can help us find a cause, see below). Since A and B are ideal terms we do not attempt to tinker with them and overcome the contradiction. Contradiction is not the same as falsification.

However, in the real world, where A and B become swans and white and so refer to actual objects, at least one of the terms could be misapplied to the real world object in question. Therefore, we can at least consider the possibility that only *one* of the propositions is falsified by the contradiction. This is what Popper does without making this step explicit. He then states that it is clear which proposition is wrong, so clear that the alternative is usually dismissed as merely 'playing with words' (Thouless 1974). But this clarity is, like induction, actually only a habit of mind as well. In the example, Popper proposes that we change the definition of swan to include the possibility that some swans are black, and does not even bother to argue against the alternative. The other way out of the contradiction is equally *logical* (even if it appears implausible because of our habit of mind). We could change the definition of black to

exclude the possibility of being applied to swans. Thus the thing that looks like a swan is actually not because it is black. The choice is between changing our definition of swan or of black. In this example we tend to prefer changing the definition of the least familiar term, and swan is a much less general term than black. In fact the same appears true in every example of 'falsification'. What seems like a logical argument for falsification could actually be an appeal to the same non-logical phenomenon of familiarity that underlies induction, and therefore causation. Put like this the notion of causation sounds, and indeed is, difficult to justify. What are the alternatives?

What if outcomes were not caused?

Another possibility to be recognised and examined is that the concept of causation, on which the apparent pre-eminence of experimental methods rests, is an illusion. Effects cannot be deduced from observing causes, nor causes from observing effects (seeing a light bulb going off does not, by itself allow the observer to deduce whether it has been switched, whether there is power failure or the bulb is broken for example, Salmon 1998) It is even possible to imagine and describe social life, and events more generally, without reference to causes. Since this is so, and we cannot see, smell, hear, measure or register causes directly it may be unwise to assume that they exist. In fact, an argument could be advanced that this is the most parsimonious, and therefore the most scientific, explanation of our observations.

A perfectly plausible alternative is one based purely on random events. A large table of pseudo-random numbers can contain arithmetic sequences, and passages of repetition, without us denying their essential randomness. The sequence '0 1 2 3 4 5 6 7 8 9' is as likely to be generated randomly as any other sequence of ten digits, such as '3 2 7 5 8 8 4 5 1 9'. Both are equally 'random' in the sense that we mean when describing such tables. In the same way perhaps the apparent regularities and repetitions that we observe more generally would be expected in a large (possibly infinitely large) universe. On this, admittedly rather extreme view, all scientific propositions are like the superstitions of a gambler who believes that stroking a rabbit foot improves their odds, or of a pigeon in a Skinner box repeating pointless actions in face of an accidental reinforcement schedule.

However, this view, while intellectually coherent, means the end of scientific endeavour and, by definition, is not one that can be logically espoused by anyone engaged in research on teaching and learning. Similarly, an economist believing that market indicators were actually following a 'random walk' could not earn a living as a predictor of these indicators, except as a charlatan.

Nevertheless, causes are seen by some respected commentators as pre-scientific. Pearson (in Goldthorpe 2001) as early as 1892 was calling the idea of causes a 'mere fetish', which was holding up the advance of correlational techniques in statistics. Russell (in McKim and Turner 1997) argued in 1968 that physics no longer seeks causes as they simply do not exist. Causality is a relic of a bygone age, like the theory that infections were caused by demons invading the body perhaps. The best we can apparently hope for is the identification of 'relatively invariant functional relationships among measurable properties'. So Russell, like Pearson, would argue that scientific laws are idealised correlations. Mathematical statements or systems of equations can describe systems but they cannot express either intention or causality. If we drop a ball in a round bowl it will come to rest in the centre. We may predict this, and say that this was 'caused' by gravity, but we can see neither the cause nor the gravity, and the cause could not be expressed mathematically. This becomes clearer if we drop two balls in the bowl.

We can model the final resting places of both balls mathematically, but we cannot use this to decide which ball is 'causing' the other to be displaced from the centre of the bowl. The events are mutually determined and this system of mutual determination is what the equations express (Garrison 1993).

In economics as well as physics some commentators have moved away from causal explanations. Wages and interest rates might be inversely related over time, but rather than deciding that one causes the other it might be more realistic to describe them as mutually determining. Mathematics (including statistics) is like formal logic (see above). It can be used to show that systems are, or are not, in equilibrium, and to predict the actual change in the value of one variable(s) if another variable(s) is changed. However, this prediction works both ways. If  $y=f(x)$  then there will be a complementary function such that  $x=f'(y)$ . Which variable is the dependent one (on the left-hand, predicted side) is purely arbitrary. Nothing in mathematics can overcome this. Nevertheless, non-causal mutuality (or concomitance) could be a perfectly reasonable and reasonably useful interpretation of many such sets of events.

What if cause and non-cause co-exist?

Another position worthy of consideration in relation to the existence of causes is that they exist alongside non-caused events. One version of this stance was taken by those advancing the teleological argument for the existence of a god. Their argument was that everything has a cause, so it is possible to follow the causal chain back to the first cause which was, for the want of a better term, god. Ignoring the simple counter-argument that the existence of a first cause actually refutes the first premise (i.e. everything has a cause), it is clear that such advocates are allowing both causes and non-caused phenomena to exist in the same universe. The same approach is now followed by economists who present evidence for rational choices as a causing agent. These choices, such as those involved in human capital theory, do not appear to work for individuals but only at aggregated levels. One interpretation therefore is that individuals operate using idiosyncratic processes that only appear to be rational when grouped. More overtly, this position was taken in the twentieth century by physicists and others believing that events at some levels are random (uncertain) while at higher levels of analysis they are patterned. In social science this belief appears in models, both quantitative and qualitative, in which the predictable components of behaviour are seen as causal in nature, and the unpredicted (and unpredictable) parts are seen as random error terms or individual whimsy (Pötter and Blossfeld 2001).

An alternative view is that all of these positions, while logically possible, are currently as invalid for the practising social scientist as the model of entirely random events. The number of potential explanations for any finite set of observations is actually infinite (created by simply adding more and more redundant clauses to a proposition for example). We overcome this practical problem, and foster cumulation, by concentrating only on the simplest explanations available. These are the most parsimonious, seeking to explain the observations we make without using additional propositions for which there is not already evidence. They are also the easiest to test, and to falsify in the Popperian model. We have no direct evidence for either causes or random events (Arjas 2001), so to use either one of them in an explanation involves making an assumption. To explain a set of observations using both involves making *two* assumptions, and is therefore unparsimonious. We have enough trouble establishing whether causes exist or not. To allow them to exist alongside unrelated phenomena makes most social scientific propositions completely untestable (for the falsification of a purported cause can always be gainsaid by

the 'whimsy' element). Perhaps this is why the social science of education shows so little progress over time.

Uncertainty could also be merely unpredictability, and it would be arrogant to assume that if we cannot yet predict a set of events then there is no more predicting to be done. Chaos theory is clearly causal but it allows for unpredictability due to complications in computation from the initial states (Gleick 1988). This unpredictability could stem from our inability to predict causatory events, or from our misunderstanding of the basic randomness of events (see above). Both explanations are plausible, but currently untestable. Using both processes together is unnecessary, and trying to combine them into one description often leads to logical difficulties anyway. For example if sub-atomic events are really random, but have an effect on larger processes which are themselves causes, then following the causal chain argument the larger 'causes' are themselves randomly determined and therefore random. And if 'random' events can have a cause then they are not random, by definition.

A more complex solution is to construct a model that involves both causation and other competing explanations of a non-determinist nature, such as intentionality through personal choice. Gambetta (1987) describes educational decisions, for example, as a product of what is available to the individual, what the individual wants and, indirectly, the social conditions which shape the individuals' intentions. However, explanations such as this are unparsimonious on two counts. Firstly, within the causal model, if a cause can be either direct or indirect, an infinite number of possible intermediate steps can always be created between the observed direct 'cause' and the hypothesised indirect one (Blalock 1964). If either explanation fits the data, the simplest solution is the best and the simplest solution cannot be both solutions at the same time. Similarly, the problem with causation is not that there are events that it cannot explain, but that it is impossible to measure. Therefore, there is no value in mixing it up with a model of intention which is also perfectly capable of explaining decisions by itself but which is also not open to observation by social scientists. Given that there is no way of deciding between them empirically, either causation or intention can be adopted (it makes little practical difference which at this stage). There is no empirical justification for working with both at the same time (any more than there is for working with causation and randomness). Rather, in a causal explanation, an intention or an individual choice can be an outcome (of social or family background for example) as well as a cause. The argument is actually about the nature of the cause (or effect), not about whether it is a cause.

#### Notions of causality

We have perhaps, as shown above, excessive confidence in the notion of causation in social science, but if we question it what is left? At one extreme if the events we observe in our fieldwork are random, with the apparent patterns appearing by chance in an infinite universe or through self-delusion like the figures observed by ancients in the night sky, then there is no social science (and no need for research). We would have to return our grant and RAE-derived funding to the tax-payer.

Even allowing for the existence of genuinely random events alongside cause and effect produces problems. If the two sets of events do not interact, then our explanation is unparsimonious (why not three types of mutually exclusive events, or thirty?). If the types do interact then randomness 'trumps' causation. A random event cannot be caused in any meaningful sense, and an event caused randomly is random (and we are back to the first extreme). One conclusion is therefore that these unobservable causes are not necessary in philosophical terms, but that they are fundamental to social science, and to

learning and understanding more generally. It follows from that, if accepted, that most debate is about what the nature of causation is rather than *whether* it is. When psychologists argue the nature/nurture controversy, or sociologists debate the relative importance of structure and agency, for example, they are simply arguing about what the relevant causes are.

Excluding a middle-way on this issue leaves two general approaches. Events are not caused (random, unpredictable and inexplicable) and the apparent regularities are due to chance. Events are caused (determined, potentially predictable and explicable) and the apparent exceptions are due to lack of knowledge. The first of these has been covered above. The second allows several different interpretations, and it is important to recognise and examine some of these as well so that we can be clearer when discussing/implying causality in our research which of these interpretations is in operation. The remainder of this section outlines a variety of characteristics for causal models, and suggests the current level of agreement about them.

One way of viewing causation is as a stable association between two elements. Where one is present the other is also, and when one is absent the other is also. It is the constant conjunction that suggests that all possible futures will be like all pasts (Hume 1962). This view of causation has two main problems: we know that it opens us to superstition, and it does not allow for intermittent association (see above). Skinner's accidental reinforcement schedule is a powerful reminder of the dangers of allowing causal models to be based only on association. Skinner's intermittent reinforcement schedule shows us how difficult it might be to shake such causal models once they have been accepted.

We can be easily fooled by association (hence the common caveats about correlations in standard textbooks), especially where these associations involve large numbers and are backed by expertise or apparent authority (Brighton 2000). A case in point appeared in one of the first Programme seminars for the ESRC Teaching and Learning Programme, where the leaders of two projects made the same argument. They accepted that correlation was not the same as causation, but suggested that multi-level multi-variable linear regression *was* able to detect causes. But linear regression however complex is still based on correlation, having all of the same limitations with the added disadvantages of being harder to understand. A similar point was made recently by Johnson (2001) about the false distinction in the US between 'causal-comparative' studies using analysis of variance techniques, and 'correlational' studies. Even though comparative models involve comparison between two or more groups (and like correlational techniques are becoming increasingly complex), they do not provide positive evidence of causation in non-experimental designs. It is, perhaps, simply their complexity and the apparent authority of the statisticians who understand them that makes others prepared to accept this falsehood.

Despite all of these caveats, purported causal models based only on association appear throughout the research literature, sometimes dominating entire fields of endeavour. Where economists talk about causation they often mean something much weaker, like Granger-causation or temporal relationships, which takes the *post hoc ergo propter hoc* fallacy of logic and converts it via a little flourish and an empirical test of 'causality' to a seemingly respectable principle. Granger-causation in economics assumes that we are working with a universe of information. If a variable is eliminated from this universal model, and this produces no change in a second variable then the first variable cannot be the cause of the second (Hendry and Mizon 1999). Otherwise it can be said to 'Granger-cause' it. The practical problem with this empirical approach to causation is that a Granger-cause and a cause are not the same

thing but they sound confusingly similar, and anyway no one actually works in the 'universe of information'. Economists use regression models very far from universal in nature, sometimes even bivariate, and still claim Granger-causation which becomes, in essence, a fancy term for a correlation. A similar approach is sometimes used in partialling variance in school effectiveness work. Here the argument is for robust dependence. A variable is not a cause if its influence (regression coefficient) is eliminated by the addition of new variables to the system. But this is clearly nonsense (Goldthorpe 2001). A causal path analysis may show that education leads to a higher income but this is very far from showing that education causes income. Robust dependence is not enough. Only a prediction from theory, or a test via intervention, can take us any further than a purely descriptive mathematical relationship.

Given the difficulty of identifying causes, perhaps the best that can be hoped for is to identify only weak causes or 'determinants'. These could be the producers of the observed effects, or they could be simply the indicators, or sign posts, of a future outcome. In fact, social scientists outside structural equation modelling use many forms of determination which are not strictly causal, including historical and structural analyses (Pötter and Blossfeld 2001). We should also accept a causal model which is probabilistic rather than deterministic in nature (Goldthorpe 2001), although we would be unable to decide whether this worked because the world is actually non-determinist, or because it is too complicated to explain fully and so we allow for error. 'The teacher may give what appears to be the same lesson in exactly the same way in a second classroom, but the outcome of the second lesson may be quite different because some un-noted variables of the setting, or the class, or the individuals within the class, are sufficiently different to affect the outcomes' (Bassey 2001, p.7). However, rather than see this as something peculiar to education or social science we should recognise that this is the common form of causal modelling. Simple deterministic causation is rare in reality, where even physical 'laws' are actually generalisations from many differing observations, or *ceteris paribus* (Hammersley 2001). Water tends to boil at 100 degrees centigrade, but it depends on the atmosphere and the purity of water. Few readers will have actually witnessed water boiling at precisely 100 degrees. The best we will have done is observed a tendency for what we call water to boil at near that point, and to have created a list of exceptions and conditions. Even such a simple law *appears* probabilistic, rather more than like the constant conjunction of strict determinism.

'It can be said to be axiomatic to any notion of causality that it only acts forward, that is, a cause must precede its effects in time' (Arjas 2001, p.60). In research, as in life, an easy assumption is sometimes made about the direction of causation that does not really stand up to scrutiny. This assumption is that one event can only be considered a cause of another if it occurs first, therefore if two variables are related then their temporal sequence defines which is the cause. For example, an analysis by Dolton et al. (1994) of data from the longitudinal Youth Cohort Study explained the labour market position of each participant in terms of their position in previous sweeps. Similarly, Gershuny and Marsh (1994) explained each participants' current employment position, as described by employment status, sector and occupational level, in terms of two main determinants, both preceding the current employment position - their initial characteristics and the accumulation of previous employment. They therefore adopted a 'recursive determination model', which was based on the causal chain approach advocated by Blau and Duncan (1967). Technical literature generally suggests methods for analysis of event histories, such as Proportional Hazard modelling, which try and explain occurrence variables in terms of prior events. In fact, a majority of studies use these unidirectional recursive models, such that the

predictors or independent variables are themselves unaffected by the outcomes or dependent variables (Berry 1984). The approach was summed up in one study thus, 'what we do now becomes what we are, and what we are in part determines what we do next' (Gershuny and Marsh 1994, p. 69). In their analysis of the determinants of unemployment, variables were entered into the model in the order that they occurred historically, from parents' occupational class through initial education to the work details. The 'effect' of the earlier episodes was assumed to be present throughout the analysis but was found to diminish over time. In this way, the past is seen as affecting the present while both can affect the future, but the future cannot affect the present and the present cannot affect the past.

However, in many respects the assumption of unidirectional causation is unrealistic (Berry 1984). Causality is merely assumed to be time-determined (Hume 1962). The relationships between data which are seemingly in a temporal sequence are often reciprocal (Hagenaars 1990). Many educational decisions may be made in the expectation of a fairly remote future outcome, for example a decision to stay on at school at age 16 in order to become a barrister. Even as early a stage as picking courses of study for GCSE can be dependent on final career intentions (Roker 1991). Rational choices can allow people to jump towards attractive options rather than being 'pushed from behind' (Gambetta 1987). In addition, the direction of the arrow of causation is not at all clear even in well-established links between variables. For example, does a higher family income lead to a better education for the children, or can stress laid on future educational plans also lead to a need for a higher level of income? Does greater investment in training lead to company growth, or are richer companies more likely to spend money on training? Analyses using only a single equation model, such that A is predicted by B and C separately assume no relationship between the predictors. Regression models are now becoming more complex, with multi-equation rather than single equation models, which allow the predictors to influence each other, so that although A may be caused by B, both A and B may be caused by C. However, in theory at least, if two variables can be reciprocally related or even if their error terms are related, then a non-recursive model must be used, but this is seldom seen in social science outside econometrics (Berry 1984). One reason for this is that as it is not possible to deduce the direction of causation from a simple association, some non-recursive models cannot be meaningfully described.

Teleology is not a respected phenomenon, and intentions are not usually seen as being causes from the future. An opinion poll of researchers would probably discover very little support for the notion of 'backwards' causation. However, a variety of situations and puzzles have been devised which expose how common the teleological explanation of events actually is.

A simple example of our habit of using backwards causation is as follows. You are appearing in a TV quiz, and are presented with three closed boxes. One box contains a prize and the other two are empty. You are allowed a free guess. If you pick the box with the prize, you win it. You select one of the boxes (box A for example). The compere, who knows the contents of each box, then deliberately opens an empty box (box C for example) and shows it to you. The compere then gives you a chance to change your mind. Do you now have any reason to pick another box (box B in this example) or to stick with your original choice? Put another way, what have you learnt from the opening of box C?

Many readers will argue that they have no reason to change their mind, but that they now have an improved chance of winning whether they stick or pick box B. People tend to claim that whereas they had started with odds of 1 in 3, they now face odds of 1 in 2. But even being tempted by this 'analysis'

displays a belief in backwards causation. Nothing that the compere has done in opening the box can change the position of the prize or, therefore, the odds of winning. When the game started you had odds of success of 1 in 3 (with box A). The prize was twice as likely to be in one of the other two boxes, even though one of the other two boxes must be empty. The fact that you now know which of the other two boxes is empty changes little. The prize is still one third likely to be in box A, and two thirds likely to be in one of the other two (which is simplified now to box B). Picking box B is twice as likely to be successful as picking box A. To consider otherwise implies that opening box C can have an effect on the actual position of the prize (cf. the problem of Schrödinger's cat).

Since the entire notion of causation has no solid evidence-base, but is chiefly a habit of mind according to Hume and others, the fact that reverse time causation is also a habit of mind gives it a very similar philosophical and scientific status as the more usual causal models. In fact, in a full determinist model of events it makes as much sense for time to run backwards as forwards (it would, presumably, not be possible to tell the difference anyway).

Nevertheless, in evaluating whether a possible theory makes sense, De Vaus (2001) suggests in addition to explaining the co-variation and time sequence, and being plausible, that the proposed dependent variable must be capable of change. While the sex of the student could affect the outcome of an assessment, the reverse could not be true. Sex would be unchanged by the assessment. In fact, we can go further than saying the dependent variable must be capable of change. It must be able to be changed by the independent variable. If there is a relationship between the level of poverty among sixteen-year-olds and their examination results, then the only causal model that makes sense in the short-term is one where poverty leads to examination results.

A possible characteristic of a good causal model is an explanatory process or theory that takes these restrictions on plausibility into account. If causation is a generative process then something must be added to the statistical association between an intervention and an outcome for the model to be convincing. The cause must be tied to some process that generates the effect. The standard example is the clear relationship between smoking and lung cancer. The statistical conjunction and the observations from laboratory trials were elucidated by the isolation of carcinogens in the smoke, the pathological evidence from diseased lungs and so on. From this complex interplay of studies and datasets emerges an explanatory theory - the kind of theory that generates further testable propositions. This is the key role for theory-building in educational research.

This brings us back to the role of experiments. Another way of viewing causation is via the effect of an intervention. If causes are not susceptible to direct observation, but what they 'cause' is effects, then at least those effects must be observable. We therefore follow the principle of 'no causation without manipulation', and attempt to mimic the classic Fisher experiment. This is the approach used by Pavlov in so far as classical conditioning involved a causal model of learning and extinction. Koch used a very similar approach of intervening and treatment removal to show causation in infections (Cox and Wermuth 2001). Unfortunately in a social science where the subject of study is people we cannot usually expose the same people both to the treatment and not, as might be possible by using two near identical cases in Physics for example. We therefore use statistical approaches (including random allocation to groups) to overcome this limitation. And this, of course, may be why probabilistic models

of causation emerge. They may reflect, not the reality of the study, but the limitation of our experimental design.

These same statistical procedures are now more widely used where an intervention is not even attempted but is replaced by further statistical controls such as weighting. There remains fundamental disagreement over the validity of these approaches (McKim and Turner 1997). Prediction, based on correlation alone, does not depend on a causal relationship, nor does it necessarily exhibit causation. This is true however impressive the prediction is - we may accurately predict the severity of a fire from the number of fire engines attending without attributing the cause of the fire to the engines (De Vaus 2001). Day always precedes night and so could explain 100% of the variance in a regression model, but that would not make it the cause. What we need to do is be creative in the consideration of alternative explanations (and then test these if practical). In fact, it is common to encounter the 'fallacy of affirming the consequent' in social science. The fallacy argues that if A is true then B will follow. Then if B appears it means that A is true. While seductive there is no logic to this argument unless it starts more strongly with 'only if'. Otherwise exactly the same argument can be made with Z (or anything else) substituted for A.

It is interesting also to consider the legal position of causation. Evidence for causation has been presented in many legal cases - a common theme in occupational medicine for example (Rom 1992). Bradford-Hill's criteria for identifying causation are widely applied (Bradford-Hill 1996). These are a temporal relationship, specificity, biological [i.e. mechanical] plausibility, and coherence. When these have been put to the test in law, the US Supreme Court has ruled (e.g. in *Daubert versus Merrill Dow*, 1993) that experts seeking to prove causation (of the toxic effect of a chemical) have to establish a greater than 50% probability based on common ground, generally agreed techniques and evidence from peer-reviewed publications. The expert does not therefore need to rely on scientific proof or certainty, or even intervention trials necessarily. Helpful, but not all essential, characteristics of the purported causal relationship are:

- confirmation of association in different studies, researchers, populations, and methods;
- frequency of association compared to frequency of either alone;
- exposure to the factor (for the individual or the population) before the onset of the disease;
- predictive relationship between the factor and frequency (biological gradient);
- isolation of factor and use as intervention to create disease;
- coherence with previous knowledge, and plausibility;
- workable previously agreed analogy;
- reduction in disease after removal of factor.

## Conclusion

Causes are particularly relevant in a climate of evidence-informed policy-making and practice for at least two reasons. Causes are really only susceptible to *testing* by intervening and measuring, the technique of randomised controlled trials and related designs which form the basis for the Campbell/Cochrane syntheses. In addition, in order to determine what works in any given situation the intervention must be proposed first (for there are an infinite number of potential interventions). While this creative phase of a study can be, and has been, inspired or serendipitous, the closest we have to a technique for generating such ideas is to try and understand why things work. This is the role of theory - not banner-waving grand theory, but attempts to provide explanations for observed phenomena in ways

that are fruitful and actually testable. A useful causal theory would have the characteristics of all of the models proposed above. It would involve conjunction (relatively stable association), a measurable effect from the intervention, and at least a tentative theoretical explanation. However, the more standard notions (Pötter and Blossfeld 2001) of cause and effect having spatial and temporal contiguity, constant conjunction, and temporal succession have all been brought into doubt in this discussion. Causation is a difficult concept for use in social science for a variety of reasons. It is difficult to define, requiring some form of agency or productive force to be meaningful, although this force is never observed and merely induced from patterns of ordered events (Blalock 1964). Some models may allow causes to operate at a distance, or even require causes to come after effects, and some may wish causes to be only probabilistically associated with effects. All of these are possible, and none of them disturb the conclusions just drawn.

Having resolved this, in practical terms cause/effect is still difficult to isolate. Given the design bias, and sampling and measurement errors in all our work we may end up with estimates, catalysts, determinants, or even 'fuzzy generalisations' rather than simple, almost mechanical, cause and effect models. While perhaps disappointing to some, this is actually inevitable. Our role as researchers is to minimise the bias and the sampling and measurement errors. Statistics, as popularly conceived, can only help with the least important of these - the sampling error (and while statistical procedures describe ideal situations, social scientists conduct their studies in, and are influenced by, real-life social settings, Gephart 1988). Overcoming the rest of the error, the bulk of it in any design, is to do with rigour. Rigour would transcend any specific approach or method. It is certainly not the prerogative of experiments (whose importance lies in the intervention only). The current paucity of experiments in social science is therefore not an excuse to evade the need for rigour. The same situation is faced in many fields such as archaeology, palaeontology and astronomy, and for more solidly practical reasons perhaps. Even cutting-edge sciences such as molecular genetics use relatively few genuine experimental designs (although the routine benchwork creates controls as a matter of course). The same situation applies in a range of scientific and quasi-scientific settings (Collins and Pinch 1993). 'Physics envy' among social scientists is misplaced, and there remain many useful strategies of a non-experimental nature that enable us to increase our confidence in perceived causal relationships (such as selection modelling, or longitudinal studies combined with triangulation of methods, see Johnson 2001).

We therefore give tentative support to increased use of intervention studies in education. The hurdles they face, apart from opposition to their scientific nature (see below) include separating the real effect of the intervention from the preferential funding that usually accompanies it, from the impact of volunteers and willing users, and other problems of theorising, generalising and scaling up. This is part of what research capacity-building is intended to address.

### **What is research capacity-building?**

The term 'capacity' may not be the best choice of a term to describe what we seek to build, usually referring to size and volume rather than 'potential' (McIntyre and McIntyre 2000) which is how we use the term here. Perhaps 'capability' is a better descriptor? It refers to what could be done 'now', or in the immediate future, in research (rather than given infinite time and resources for example). This involves a variety of factors including expertise, motivation and opportunities. It clearly also requires a

widening of the methodological approaches considered, and therefore possible, for any piece of research – so that in the future if a new researcher rejects the use of a field trial it is because such an approach is not appropriate to the question, not because they could not conduct one, did not know anyone who could conduct one, or did not even consider the possibility.

RCB is not, or should not be, simply a technical exercise about methods of data collection or analysis. 'Persuasiveness [of findings] may require more than simply strong research design... the potential for research to contribute to practice depends on its ability to influence teachers' thinking' (Kennedy 1997, p.7). In order to be effective, social science knowledge must be appropriately packaged and mediated by practitioners so that they can 'make it their own'. It is therefore essential that researchers, policy makers and practitioners develop new ways of working together. This new approach could start with, but not be confined to the issues of research quality, impact, focus and continuity identified in NERPP (1999) - and these categories provide a useful basis around which to outline our proposals for research capacity building.

*Focus:* Public educational research has been criticised as being 'mile wide but inch deep', providing bits and pieces of evidence in many loosely related areas. Research in the US is redesigning itself around a model of focused effort on a critical problem-centred agenda, fostering a high standard of quality and rigour, and collaboration between research and practice (NERPP 1999). It is this model that also applies in the UK to the Teaching and Learning Programme, as a relatively long-term commitment of public money to a set of tightly identified key issues.

*Continuity:* The same reports in the US have called for greater continuity in research pursuits, reaching across fashions in methods or topics and across political administrations, thus not being tied to political timetables. Researchers need time for reflection and impact, and long term independence from (despite sensitivity to) the current political agenda (NERPP 2000). Findings should be implemented via greater use of research syntheses (also using techniques such as meta-analysis) and greater collaboration between researchers, and between researchers and other professionals in education. To some extent, achievement of this political determination is beyond the Teaching and Learning Programme (in fact, the UK *may* be moving in the opposite direction towards a model of greater political influence on research outcomes). Nevertheless, its eight year period, and commitment to greater use of large-scale work, promises a move in this direction.

*Quality:* Although the relevance of educational research has been called into question (Hillage et al. 1998), it is generally issues of quality that have attracted greater attention (Tooley and Darby 1998) and that have been used to provide pressure for greater political influence. Strategies for packaging of results and dissemination, and therefore for the successful use of research findings, are bound to fail if those findings are deemed somehow not trustworthy. Genuine improvements in practice and policy are more likely to be based on good social science than on 'craft principles'. Good social science will generally reflect scientific principles and rigorous standards, and share scientific norms such as explicit hypotheses, sound designs, appropriate measures, quality data, and logical analyses (NERPP 2000). Long term then, these are also likely to be the criteria for believable and useable results (rather than what Bridges 1999 contrasts as 'dull nonsense').

*Impact:* A key question is ‘how can the use of research knowledge be increased in schools and school districts?’ (NRC 1999, p.2). As an academic community we may have several excuses for the difficulties and complexities encountered in our research, but we have fewer for our weaknesses in converting our findings into useable formats. Probably no other public sphere rests on such a slight research base, with personal experience and ideology so commonly used in policy formation (NRC 1999). Research impact stems partly from the quality and therefore the believability of the findings, and partly from an increase in the desire and willingness of practitioners to use research as a basis for professional change. What is needed is a two-way conduit in which researchers reach out to educators, providing credible evidence presented understandably, but in which policy and practice also has a responsibility to seek answers in good research. It is hoped that the National Educational Research Forum (among other recent developments) will take a lead in this.

The Teaching and Learning Research Capacity-Building Network is based at the Cardiff University School of Social Sciences. It comprises expertise in the substantive areas of education, training, health, psychology, social work, science, justice, labour markets, industry, regional planning, human geography, sociology, social policy, criminology, economics, and politics. It is acknowledged as a centre for the emergent ‘new political arithmetic’ in social science. We are especially fortunate in having a number of research capacity centres already in the School, most notably the Health and Social Care Research Support Unit, and the research focus on Community-based Studies of Health Education and Promotion. This interdisciplinary strength in breadth allows us to cross-fertilise ideas between subject areas (for example using our extensive experience in health trials for educational research development). It also gives us considerable experience in the use of large-scale ‘quantitative’ methods, combining these with qualitative methods, and the building of theories based on empirical evidence arising from a combination of methodological approaches. In particular, we have an international track record in the conduct of ‘large-scale, quantitative research, including experimental/quasi-experimental designs’. These skills can now be applied to assist in the development of ‘what works’ studies of the effects of different approaches to teaching and learning across different contexts.

The Executive Group includes some of those who have been advocating a change in approach when researching education to include a greater range of evidence, more rigour but also greater caution in drawing conclusions (e.g. Gorard et al. 1999, Gorard 2001c), those involved in randomised controlled trials and multi-level modelling (e.g. Moore 1996, Moore et al. 2000), other areas of large-scale data analysis and statistical modelling (e.g. Prandy and Bottero 2000), and diverse substantive areas of education and training research (e.g. Crozier 1997, Furlong 1996, Rees et al. 2000). The wider Network will include representatives from the projects funded as part of the Programme, particularly to enrich its existing expertise in theory-building relating to teaching and learning for different age groups and pedagogical settings. It will also include a large number of other research staff from the School of Social Sciences, expert consultants from other institutions where applicable, practitioners and policy-makers, representatives from the ESRC, members of the Programme Steering Committee, and international advisers.

## **What we propose?**

Our key objective is to co-ordinate the research capacity building activities of the Teaching and Learning Research Programme, and help produce a significant enhancement in the methodological skills and approaches of a substantial body of UK educational researchers, both within the Programme and beyond. Our particular emphases are on rigour in the conception and conduct of research, and on consequent impact in practice. In this context, discovering approaches currently in use that are less beneficial could be as important as finding new approaches to teaching and learning that are more so. The activities undertaken in pursuit of these objectives include a skills and needs audit for educational research, the subsequent evaluation, brokering and provision of relevant training materials, and the production of recommendations for further capacity building exercises (including the development of new ideas for training at this level) based primarily on a terminal evaluation of our own work. As a result of the skills and needs audit we will generate a 'skills' profile for each researcher within the Programme, a map of the 'needs' of the UK educational research community, and a corresponding suggested development plan. While we recognise the potential difficulties of the task, especially given the fact that our work starts only when the substantive projects have already been selected and their researchers appointed, we also expect to be part of a wider movement that will produce a better balanced educational research community in the UK, with significantly greater collective skills in conducting and using the results of large-scale studies (ideally in tandem with other data sources). We aim to produce an increased interest in a wider range of methods than is standard in UK educational research today, among Programme researchers, the research community, and other consumers of research evidence.

Therefore, our attention will be on an up-and-coming generation of educational researchers, perhaps more so than on the more established project leaders within the Programme. For this latter group we would like to assist their greater use of inter-disciplinary theories and methods, enhancement of the training opportunities available at a high level, and develop models of the transformation of knowledge through to its embodiment in practices which could raise learner attainment. None of these objectives is incompatible with our intention also to provide or identify suitable training for all levels of need within the Programme (and for other researchers in mid-career), and to be responsive to all requests for advice and support from these. We do feel that it will also be necessary to conduct 'why we need...?' courses both for the investigators in the Programme and for the research community more widely to help explain the purpose and potential benefits of research capacity building and of the range of the activities we propose conducting. Where such courses are held early on in the exercise, then feedback from them can be taken into account in planning the remainder of the work programme.

*Development of skills in the design, conduct and management of quantitative studies.*

It is expected that this will be highlighted again by the audit of research skills and needs as a priority area. It is therefore likely that specific training courses will be run on research design, secondary analysis, design and analysis of quasi-experimental and experimental studies, and on the conduct and management of large scale studies including trials. Since high quality experimental research studies are expensive to undertake, it is crucial that the interventions they are designed to evaluate are high quality and underpinned by theory. Investigators must explicitly recognise the assumptions underpinning the intervention and the postulated mechanisms and processes by which the desired outcome will be achieved. Consequently, appropriate outcomes and process measures can be incorporated into study design. The Network is able to draw on developments from allied disciplines in social and health sciences to inform theory-building in teaching and learning.

The greater use of quasi-experimental designs, or randomised controlled trials (usually via natural clusters) is one area of promise for the encouragement of good social science in educational research. This is increasingly recognised as the gold standard in clinical research (Featherstone and Donovan 1998), and as an ideal to which educational research should aspire. However, the application of these designs in practice is not unproblematic, neither in the evaluation of complex health services interventions (MRC, 2000) nor in the field of education (Hakuta 2000). There are differences between education and the models of research and development used in industry and biomedicine (the inability to legislate for teacher classroom behaviour being one). It has been argued that outcome measures in medical matters are generally less problematic than in education (for example the measurement of 'lifespan' may be less ambiguous than that of 'standards'). It is also the case that the power of the experiment comes not from the design alone but from the power of the questions to which experiments *can* be addressed. Such designs should therefore be additional to, not replacements for, other recognised modes such as detailed case studies and secondary analysis (and it should be noted that multiple perspectives and approaches are used relatively unproblematically in natural sciences).

*Articulation/combination of qualitative approaches with quantitative studies.*

There is little doubt that the best evidence on effectiveness of teaching and learning practice will be accrued by the integration of qualitative methods within high quality research designs with quantitative outcomes. In health services research, the recent MRC document on the evaluation of complex interventions (MRC 2000) highlights the importance of qualitative methods in the development and formative evaluation of new interventions, and also the essential contribution of qualitative methods to process evaluation within trials, and to identify and measure unexpected impacts. Most of the problems in combining data sources are essentially practical ones, but their solution is hampered by those who oppose the whole idea of sullyng their mono-method approach.

It is particularly important for the well-being of educational research that we do not waste time in methodological paradigm wars (as distinct from spending time on the development of all methods, see Mahoney 2000). In particular we need overcome the false dualism of 'quantitative' and 'qualitative' approaches (as argued recently by Pring 2000 for example). The supposed distinction between qualitative and quantitative evidence (Popkewitz 1984) is essentially a distinction between the traditional methods for their analysis rather than between a philosophy, paradigm, or method of data collection (Frazer 1995). As Heraclitus has written, 'logic is universal even if most people behave differently' (for if logic were not universal we could not debate, so making research pointless). It would be difficult to sustain the argument that all methods, including data collection, carry epistemological or ontological commitments (Bryman 2001). Most crucially for the opponents of combining data, it is important to realise that 'qual' and 'quant' are not differing research paradigms (at least not in the sense that Kuhn uses the term). To some extent all methods of educational research deal with qualities, even when the observed qualities are counted. Similarly, most methods of analysis use some form of number, such as 'tend, most, some, all, none, few' and so on. This is what the patterns in qualitative analysis are based on (even where the claim is made that a case is 'unique' since *uniqueness* is, of course, a numeric description). Words can be counted, and numbers can be descriptive. Patterns are, by definition, numbers, and the things that are numbered are qualities. One practical reason would be that we could cease wasting time and energy in pointless debates about the virtues of one or other.

*Transformation of research based knowledge through to its embodiment in practices relevant to enhancing learner attainment.*

It has been suggested elsewhere that research findings do not really matter, in the sense that actors take so little note of them (Gazial and Blass 1999). The problem of the relationship between research and policy/practice is a complex one. Users do not have time to read original reports and data, but could be misled if presented with only simplistic summaries (Hammersley 2001b). The latter also denies practitioners the chance to debate and reflect on the evidence (Willinsky 2001). We recommend that the Programme as a whole sets out to work in what has been termed “Pasteur’s quadrant” (Resnick 2000), where the quality and practical relevance of research are pursued in tandem (also termed interactive social science). The old linear pipeline view of research leading to development, dissemination, evaluation and eventually to programmes of amelioration takes too long (and therefore is not working). It makes feedback difficult, but above all it means, by definition, that the nearer one is to practical realisation the further one is from the research on which the solution is based. The quality of fundamental knowledge *and* considerations of use are both important dimensions of research. Their presence or absence creates four quadrants, and both occur in Pasteur’s quadrant where work is done to improve complex systems, co-develop research and practice, continually refine solutions, build theory, while simultaneously developing the ability of solutions to ‘travel’. To a large extent, these criticisms and potential solutions apply to the travel of methodological change as well as to substantive findings. Therefore, Research Capacity Building has a key role in communication and improving communication between *all* parties as well as in furthering the professional development of some of them.

However, we should not exaggerate the potential for the future of ‘interactive social science’ by itself. Research into adult and continuing education has traditionally had very close links between researchers and practitioners (i.e. they are the same), and has also tended to use a lot of ‘theory’. But it would be difficult to sustain the argument that this field was therefore superior in its research to others (and it might even be easier to argue the reverse). Even where policy-makers and practitioners genuinely welcome research evidence, and much research actually leads to discomfort for politicians, other factors make the link unlikely – such as the need for quick results to fit an electoral cycle (Levacic and Glatter 2001). Safe knowledge about topical issues is often simply too late to matter. These, and related problems, illustrate why we feel that the problem of transforming results needs to be attacked from both ends. We need good science and we need to prepare consumers of research concerning what is possible and what is not.

Review of the research resources available within the Programme and beyond and identification of opportunities for adding value and collaboration

It is recognised that this project will need to be pursued with sensitivity to the possible discomfort that the audit and its accompanying work plans may engender in some sections of the research community, especially in light of the considerable expertise already present in the Programme. We must recognise that some researchers nearer the end of their career may be resistant to modification of what, for them, have been successful techniques. It is also the case that we cannot, and would not wish to, ignore the new ethical issues that may arise in the extensive use of experimental designs in education for example. Research protocols cannot be simple ideals or mindless ‘recipes’, but practical usable approaches to gaining more cumulating knowledge of what works than hitherto. All 14 projects have been commissioned in the face of considerable competition because they can already provide the research

skills necessary to meet their objectives. We are therefore in a position analogous to conducting an industrial audit based not on the existing production system but on a projected change in the method of production, and this position must be made clear at the outset and reinforced regularly. Nevertheless, there will almost certainly be those who do not share our vision of the future of educational research. It will therefore also be important to make clear that the purpose of the audit is to help blend new research approaches into the existing mix of protocols rather than seeking to replace or reprimand the latter in any way. There may be other natural difficulties arising from our focus on the 14 projects in the Programme, since despite having agreed to the importance of capacity-building there may be a reluctance on their part to devote much time to these activities at the expense of a more traditional concentration on their own 'job-in-hand'. Part of our role is to transform this view (if it is encountered), and to help them see their existing full-time objectives as well as their future career development as consistent with this transformation. In such ways we expect to be able to enthuse some individuals, and so help lead a substantial proportion of the research community towards a greater appreciation of the value, and limitations, of good social science.

The researchers in the Programme already represent a range of important and valuable research skills, and relevant knowledge of teaching and learning. To these will be added the skills of the research staff they appoint. They will have, or have made provision for gaining, the skills they require for their own projects (by definition). Our skills audit therefore has the following further aims: as a cross-project body to identify and suggest co-operative and value-adding activities based on existing skills, to 'find' skills that could be useful for other projects, and to find skills and skilled researchers who can contribute to the work programme of the Network. The skills/needs audit will therefore concentrate not on the training of researchers for each project (although we would be happy to assist where this is an efficient use of resources), but on increasing the skills-base of the Programme and research community more widely. Additionally, we will identify areas where closer co-operation will bring added value, and specific examples that we can use as case studies of the problems faced or methods of overcoming skills deficits, and examples of innovative practice justifying wider dissemination. In these ways we can help co-ordinate the capacity building already likely to be an integral part of each project.

#### Support for research capacity building activities and training within the Programme

The chief thrust of our own training and capacity building activities will be on high level research skills training. The precise nature of these courses must depend on the results of the skills and needs audit, and related steps. However, on the basis of our experience so far these activities are likely to include high level training in complex project management, design and analysis of quantitative research studies including quasi-experimental and experimental designs, the secure integration of qualitative and quantitative evidence, theories of measurement, methods of sampling, the evaluation of impacts, research ethics, new conceptual developments, expanding opportunities for inter-disciplinary approaches, and the conversion of findings into policy and practice. A key issue for this training is a careful consideration of when particular designs are appropriate for gathering knowledge about teaching and learning, and when they are not. In particular, we expect to use our experience of cluster (community) randomised trials in health services research and elsewhere to assist in the creation of appropriate protocols for experimental work in education, to help face up to the many practical, diplomatic and ethical issues involved, to help in the conduct and management of experiments, and to provide training in the use of appropriate software for analysis. We will also work on developing the current theoretical and conceptual bases for such studies. Some of this training will be not be high-level

and method-specific but more generally about research appreciation, for it is important, in our view, that high level work is both done and then understood by a wider research community.

The training programmes so created will be pro-active, flexible in delivery, and will take on board suggestions and formal feedback resulting from our early monitoring. There will be a combination of face-to-face and interactive distance/electronic delivery. Some courses will be intensive, perhaps over the summer period and at weekends. Other training development will be responsive, such as a helpdesk, email helpline, and interactive website. Some will be indirect in the form of publication of researchers 'toolkits', and the provision of a gateway to other high quality training delivery. Much contact with researchers will therefore be virtual, mixed with longer face-to-face experiences, and to provide for a series of senior visiting fellowships for Programme grant-holders. We will operate a 'buddy' system whereby one of the researchers will work closely with another, perhaps in modelling for the first time. In addition, we propose the use of 'learning sets' in which sub-groups of researchers in the Programme form cross-project small groups who meet or otherwise interact for the purposes of specific training needs identification, mutual support and intellectual stimulation. Sets such as these have been successful in similar previous ventures in medical research. This approach will also involve placements across the projects in the Programme, where we discover matching skills and needs in two or more projects. All training will be free and expenses-paid to TLRP members.

The emphasis here is on transforming the ability of the educational research community to conduct large-scale studies of teaching and learning, and to convert their results into useable and practical knowledge. However, it should be noted that just as great a transformation needs to take place among the 'consumers' of research. An appreciation of secondary analysis, high quality survey methods, and above all experimental designs will become essential for all education researchers who wish to read and assess the work of others. We cannot allow a schism to develop, for example, where some researchers use large-scale trials and others do not understand them at all. Lack of understanding could lead some commentators either to a lack of acknowledgement of the results of trials, or to their uncritical acceptance. Neither of these outcomes is desirable, so all researchers, whatever their chosen primary methods, require a working knowledge of new protocols as they emerge. The same, to a lesser extent, is true for senior practitioners and policy-makers. Similarly all researchers, whatever their chosen primary methods, can benefit from contextual analysis using official secondary data for example. Therefore, the planned activities of this capacity building Network should be aimed as researchers as 'doers', but also at researchers and others as consumers and reviewers of educational evidence.

Dissemination and training targeted at the wider research community (including active researchers in 'user'/'practitioner' communities)

Although the activities of the Network will be focused primarily on the research capacity of the Teaching and Learning Programme, once these activities are available the increased cost involved in including other researchers will be minimal. We will therefore advertise our activities and services in a variety of ways. All training will be free, but unlike members of the Programme, external users will be responsible for their own travel and subsistence expenses. In addition to this potentially unrestricted access to training, and the involvement of practitioners in the Network, we will publish reviews of leading edge conceptual, theoretical, method work, best practice guides, toolkits and other publications aimed at the research community. Those in the Programme and others in the Network will be sent all of these as a

matter of course. Similarly our 'route maps' of access to wider training resources, not just in education, will be made more generally available, via an interactive website for example.

Linked to these 'dissemination' activities is the consideration of dissemination more widely. As shown above, the links between research and practice have traditionally been poor – partly due to the quality of the evidence, partly the style and packaging, but mostly the lack of speed with which results are converted into usable programmes of action. Education is currently not a cumulative social science. We intend to have some impact on this, by trialling and publicising methods of packaging good results, including research syntheses and meta-analyses. As a by-product of these trials we will therefore be able to publish examples of the results of these approaches in areas of substantive interest to the Programme, not also covered by the other bodies with whom we will be co-operating. This work will draw on the rapidly developing area of 'implementation research' in health services research, which tries to identify how best to get research evidence into practice. However, an early lesson here has been that the quality of research studies, and the resultant evidence, needs to be high before practitioners and policy-makers will act on that evidence. The most effective tool to persuade practitioners to adopt a more evidence-based approach is to have high quality studies that have clear implications for practice, and the activities of RCB aim to help develop and support an infrastructure to produce such evidence.

We believe that both speed of dissemination, and high quality of research are crucial to the successful utilisation of research evidence in practice. We do not, however, believe that they are sufficient in themselves. The analogy of knowledge transformation suggests working via a conduit (such as a regulatory body) but also emphasises the need for motivation (i.e. there must be a practical problem for the evidence to solve), and an awareness of that need and an incentive to solve it. Evidence in isolation will not improve teaching or learning. Therefore, in addition to assisting the building of research capacity *per se*, we will also be working on new approaches to making educational research make a difference.

### **What problems do we face?**

Clearly research capacity building would not be easy to conduct or monitor even in an ideal situation. The task is a considerable one, and one that is genuinely new - meaning that there is little from the past to guide us. In addition, it is likely that the project faces specific problems, and a few of these are outlined below.

The move towards research-capacity building for existing researchers is a novel and innovative strategy, and, as we have made clear, it is one that we fully support. Nevertheless, it is likely to raise some tensions within the education research community, and these are already apparent, to some extent, in the comments from reviewers. First: this is not a standard research project, and cannot be judged or evaluated in the usual way. We need to create new methods of judging our performance, but there is still an important element of trust involved in this award. We will work with the Programme and our own Advisory Group to develop methods of monitoring and evaluating our activities, especially in terms of long-term impact. Second: there is a tension between support for existing research and helping transform future research applications. We have been directed quite explicitly towards the latter. The TL Programme Projects have already been decided, staff appointed, and in many cases fieldwork commenced. In the opinion of the ESRC these projects are already able to meet their own objectives,

and are deemed to have access to the necessary research skills. The comments of several referees, however, imply that this is still unclear to some of them. Third: there is a tension between the deficit model of educational research that lies behind the award and the need to share expertise within the Programme and wider community. The notion of 'deficit' is expressed clearly in the call for applications, and is reinforced by the McIntyre Report. There *are* 'system-wide' deficiencies in large-scale quantitative approaches. However, in order for capacity-building to be effective the exercise cannot be approached simply in those terms.

Our key contacts will be those within the TLRP, who have recently received considerable funding to conduct their own research (for many of them the largest single grant of their professional lives). Whilst they signed an undertaking to build research capacity, with our support, as part of their project, this is likely to be far from a priority for many (and they will have their own problems of the traditional type, concerning access or contracts for example). One major reason for this priority is that their main intellectual interest is in their project. Another is that unlike in former programmes, each project has to show that it has made a difference to attainment (quite a daunting task). Yet another is that projects have traditionally been assessed in terms of their research outcomes not their research capacity building, and so habit will encourage grant-holders to concentrate on the former. And, despite the innovative nature of the TLRP brief, they may be advised to do so since the reviewers are likely to have the same traditional view. Programmes have anyway sometimes been little more than the sum of their parts, and considerable vigilance and energy is required to combat this tendency.

There remains considerable confusion over the precise nature and purpose of the RCB project. This confusion (even disagreement) is evident in the Programme Steering Group, and in the comments of ESRC referees sent in response to our bid. This confusion clearly extends, to some extent, to the project researchers, and even more therefore to the other members of the TLRP (the intended users of this project). For example, while the project is clearly intended to have an impact for the future, for the nature and range of proposals submitted to the next such programme perhaps, some comments have suggested that it could be a service or support facility for the existing TLRP projects instead. The ESRC have already funded the 14 projects and agreed that they have or can hire researchers with the necessary skills to the job. RCB has no part to play in this. Again, while the project is clearly targeted at the research and skills profile of individuals, most other projects worry less about this and more about the combined skills profile of the team. Such projects might be tempted to argue that they do not wish to undertake RCB since their team is already self-sufficient. Such an argument has the same flaw as above. RCB is not for these projects (although it may have beneficial impacts in the short term), but for the future. Nevertheless the challenge of providing a programme of training and support that will appeal to the experienced grant-holders, their somewhat less experienced researchers, and to the inexperienced practitioner-researchers is considerable.

More generally, outside the TLRP, there is likely to be a feeling that RCB is simply yet another in a catalogue of initiatives apparently duplicating themselves, and getting in the way of actually doing any research. One reaction to the introduction of NERF, for example, was that it was a disciplinary project stemming from the 1996 Hargreaves lecture intending to tell researchers what and how to research (Ball 2001). Given that higher education already faces hurdles such as the CVCP, RAE, QAA, TTA, OFSTED, transparency review, ESRC recognition and so on, perhaps the charge of serious duplication of reviews of research training and skills is well-founded. We need to work with these other bodies as

far as possible to reduce this duplication. The difference between RCB and the 'audit culture' is that the first is not intended to be prescriptive. In its simplest form, researchers and users will specify what they see as gaps in the distribution of research skills, and we will offer or broker training to deal with it. In the end, of course, it is up to individual researchers to decide whether they need or want to take part.

Another, wider, problem for RCB is the practical one of recognition of its fruits. If the peer-review system continues as it is at present it is possible that good research will not be rewarded by publication, grant-income or even RAE success (and of course the following comments do not only apply to education). Journals tend to display considerable cronyism (Travers and Collins 1991) – the current special issue of the *Journal of Education Policy*, for example, contains papers only from friends of the editor who all work in two institutions in London (the one he is in and the one he recently moved from). An RAE member recently published a methodological critique of a piece of work without bothering to read the method involved first (Gorard 2001d). Such examples are cited to remind readers that if there are deficits UK educational research then it may be the 'reward' rather than the 'production' system that is at fault. Cronyism and particularism distort the review process. Although the two are not wholly distinct, this is more dangerous when it is cognitive (i.e. based on intellectual similarity) rather than institutional (which tends to be more visible). Papers are routinely rejected by their expert referees, perhaps for ideological reasons, or because they dispute the views of current 'experts' in the field (the referees). Papers and proposals are generally rejected where the reviewers disagree, since this is seen as the protection of quality. Therefore new work, radical approaches, divergent thinking, and even new techniques are likely to lead to rejection. The mechanisms of review provide short-term inequity and a longer-term encouragement of orthodoxy (Travers and Collins 1991). The worst acceptance rates are for inter-disciplinary work, of precisely the kind that RCB is advocating. Therefore the work of RCB is likely to highlight wider issues such as these.

There will, of course, be many commentators who object even to the very tentative support for evidence bases, statistical modelling and interventions expressed here. There are already comments that social science cannot be like natural science, or even that education cannot be a social science. To think otherwise is to be accused of 'scientism' and 'neo-positivism' or other unworthy and essentially non-academic terms of abuse (and it is interesting how many of these terms, such as 'positivist', are now only ever used to describe others). The supposed incompatibility of science and social science is generally based on a misconception (Hammersley 2001), usually of the nature of science. One problem here may be that science is not really what many non-scientists tend to think it is (Collins and Pinch 1993). As with ethnostatistics (see above) it is actually impossible to separate science and society. Science is an essentially anarchic, not rule-based, activity (Feyerabend 1993).

Much of the misunderstanding and distrust of science may also come as a reaction to its attempted glorification by others, such as politicians, historians and journalists. Humphry Davy 'invented' the miners' safety lamp in 1815, and this has been hailed as a hallmark achievement for science and for humanity (since it was intended to save lives). In fact, proportionately more lives may have been lost as a result since miners were then sent to workings where there was a build up of methane (which were previously considered too dangerous). This does not suggest that the lamp did not work in a technical sense (which it did), but the downside is perhaps obscured in traditional accounts of science. An unsophisticated observer might therefore deduce that since the safety lamp patently did not save lives,

then the lamp itself did not work. The scientific baby is often thrown out with the social bathwater (Chalmers 1990).

As Lakatos (1978) points out maintenance of the status quo in any scientific endeavour 'is achieved by censorship' (p.44). It is not that 'normal science' in Kuhnian terms does not exist but that perhaps it should not exist. It may be simply bad science which is uncritical rather than cumulative in nature. It is often based on actual practices that differ from those stated (i.e. there is deceit, either of the self or the audience). Normal science may therefore give an appearance of harmony, and fitting together because its practitioners conceal their actual methodological divergence in practice (Gephardt 1988). What have been termed revolutions or paradigm shifts are not therefore irrational, non-empirical or even gestalt-switches necessarily. Perhaps these would be normal science, if all researchers had the courage of a few to admit that they are not following a previously agreed recipe. The need to make new explanations consistent with some already established body of theory tends to stifle progress (Feyerabend 1993), so *some* proliferation of ideas, use of apparently obsolete notions, even inconsistency within findings or between findings and explanation can be beneficial. The church indictment of Galileo was probably more rational than his own evidence, and even Newton faced contrary 'falsifying' evidence from the outset. It is often not these factors that decide on public acceptance, and many ideas now widely accepted were originally resisted and derided. Of course, this does not mean that we should accept anything at all. It merely illustrates that science is most definitely not what many imagine it to be. It has been used as a kind of straw target for our fears. It has no universally agreed account of its nature or method (Chalmers 1999). It offers no certainties. Yet people still booked holidays to coincide successfully with the return of Halley's comet, or the eclipse of the sun. Even fully-fledged relativists take care when crossing the road (Bailey 2001).

## **Endword**

This paper attempts to take a balanced overall view of the role and function of research capacity-building. The ideas presented here all have advantages and limitations, and we have attempted to clarify these while presenting our plans. In doing so we tread a kind of middle-way, and run the danger of falling between two stools. The balance is *between* the various sections not within each of them, and for this reason we consider it important that the paper should be read in its entirety.

## References

- Adair, J. (1973) *The Human Subject*, Boston: Little, Brown and Co.
- Arjas, E. (2001) Causal analysis and statistics: a social sciences perspective, *European Sociological Review*, 17, 1, 59-64
- Ayer, A. (1972) *Russell*, London: Fontana
- Badger, D., Nursten, J., Williams, P. and Woodward, M. (2000) Should all literature reviews be systematic?, *Evaluation and Research in Education*, 14, 3&4, 220-230
- Bailey, R. (2001) Overcoming veriphobia - learning to love truth again, *British Journal of Educational Studies*, 49, 2, 159-172
- Ball, S. (2001) 'You've been NERFed!' Dumbing down the academy, *Journal of Education Policy*, 16, 3, 265-268
- Bassey, M. (2001) A solution to the problem of generalisation in educational research: fuzzy prediction, *Oxford Review of Education*, 27, 1, 5-22
- Bernard, H. R. (2000) *Social research methods: qualitative and quantitative approaches*, London: Sage
- Berry, W. (1984) *Nonrecursive Causal Models*, London: Sage
- Blalock, H. (1964) *Causal Inferences in Nonexperimental Research*, Chapel Hill: University of North Carolina Press
- Blau, P. and Duncan, O. (1967) *The American Occupational Structure*, London: John Wiley
- Bradford-Hill, A. (1966) The environment and disease: association or causation?, *Proceedings of the Royal Society of Medicine*, 58, 285
- Bridges, D. (1999) Educational research: pursuit of truth or flight of fancy?, *British Educational Research Journal*, 25, 5, 597-616
- Brighton, M. (2000) Making our measurements count, *Evaluation and Research in Education*, 14, 3&4, 124-135
- Brown, A. and Dowling, P. (1998) *Doing research/Reading research: a mode of interogation for education*, London: Falmer
- Brown, M. (1998) BERA President's letter to Secretary of State, *Research Intelligence*, 65, 14-15
- Bryman, A. (2001) *Social research methods*, Oxford: Oxford University Press
- Campbell, D. and Stanley, J. (1963) *Experimental and quasi-experimental designs for research*, Boston: Houghton Mifflin
- Chalmers, A. (1990) *Science and its fabrication*, Milton Keynes: Open University Press
- Chalmers, A. (1999) *What is this thing called science?*, Milton Keynes: Open University Press
- Clegg, F. (1992) *Simple Statistics: a course book for the social sciences*, Cambridge: Cambridge University Press
- Collins, H. (1985) *Changing order: replication and induction in scientific practice*, Chicago: University of Chicago Press
- Collins, H. and Pinch, T. (1993) *The Golem: what you should know about science*, Cambridge:
- Cox, D. and Wermuth, N. (2001) Some statistical aspects of causality, *European Sociological Review*, 17, 1, 65-74
- Cook, T. and Campbell, D. (1979) *Quasi-experimentation: design and analysis issues for field settings*, Chicago: Rand McNally
- Cox, D. and Wermuth, N. (2001) Some statistical aspects of causality, *European Sociological Review*, 17, 1, 65-74

- Crozier, R. (1997) *Individual Learners: Personality Differences in Education*, London: Routledge
- Daubert versus Merrill Dow (1993) 509 US 579, 113 S. Ct 2786
- Davis, A. (2001) Effective teaching: some contemporary mythologies, *Forum*, 43, 1, 4-10
- De Vaus, D. (2001) *Research design in social research*, London: Sage
- Dean, H. (2000) *What's the evidence for 'evidence-based' social policy? Welfare reform, low-income families and the role of social science*, presented at fifth ESRC seminar on Measuring Success: what counts is what works, Cardiff: September 2000
- Department of Health (1997) *Research and development: Towards an Evidence-based Health Service*, London: Department of Health
- Dolton, P., Makepeace, G. and Treble, J. (1994) Measuring the effects of training in the Youth Cohort Study, in McNabb, R. and Whitfield, K. (Eds.) *The Market for Training*, Aldershot: Avebury, p. 195
- Dyson, A. and Robson, E. (1999) *School, family, community: mapping school inclusion in the UK*, Leicester: Youth Work Press
- Ellmore, P. and Woehilke, P. (1998) *Twenty years of research methods employed in American Educational Research Journal, Education Researcher, and Review of Educational Research from 1978 to 1997*, (mimeo) ERIC ED 420701
- Featherstone, K. and Donovan, J. (1998) Random allocation or allocation at random? Patients' perspectives of participation in a randomised controlled trial, *British Medical Journal*, 317, 1177-1184.
- Feyerabend, P. (1993) *Against method*, London: Verso
- Field, A. and Wilkinson, L. (2001) Getting your numbers wrong, *The Psychologist*, 14, 6, 316
- Firestone, W. (1987) Meaning in method: The rhetoric of quantitative and qualitative research, *Educational Researcher*, 16, 7, p. 16
- Firestone, W.A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland, *Educational Evaluation and Policy Analysis*, 20, 2, 95-113
- Fisher, R. (1935) *The design of experiments*, Edinburgh: Oliver and Boyd
- Fitz-Gibbon, C. (1996) *Monitoring education: indicators, quality and effectiveness*, London: Cassell
- Foster, P. (1999) 'Never mind the quality, feel the impact': a methodological assessment of teacher research sponsored by the Teacher Training Agency, *British Journal of Educational Studies*, 47, 4, 380-398
- Frazer, E. (1995) What's new in the philosophy of science?, *Oxford Review of Education*, 21, 3, 267
- Furlong, J. (1996) Do Teachers Need Higher Education, in Furlong, J. and Smith, R. (Eds.) *The Role of Higher Education in Initial Teacher Education*, London: Kogan Page
- Furlong, J. (2000) Intuition and the Crisis in Teacher Professionalism, in G. Claxton and T. Atkinson, (Eds.) *The Intuitive Practitioner*, Buckingham: Open University Press
- Gambetta, D. (1987) *Were they pushed or did they jump? Individual decision mechanisms in education*, London: Cambridge University Press
- Garrison, R. (1993) Mises and his methods, pp.102-117 in Herbener, J. (Ed.) *The meaning of Ludwig von Mises: contributions in economics, sociology, epistemology, and political philosophy*, Boston: Kluwer Academic Publishers

- Gazial, H. and Blass, N. (1999) The extended school day in Israel: do research findings really matter?, *Educational Policy*, 13, 1, 166-179
- Gephart, R. (1988) *Ethnostatistics: Qualitative foundations for quantitative research*, London: Sage
- Gershuny, J. and Marsh, C. (1994) Unemployment in Work Histories, p.66, in Gallie, D., Marsh, C. and Vogler, C. (Eds.) *Social Change and the Experience of Unemployment*, Oxford: Oxford University Press
- Gleick, J. (1988) *Chaos*, London: Heinemann
- Goldstein, H., Huiqi, P., Rath, T. and Hill, N. (2000) *The use of value-added information in judging school performance*, London: Institute of Education
- Goldthorpe, J. (2001) Causation, statistics, and sociology, *European Sociological Review*, 17, 1, 1-20
- Goodman, N. (1973) *Fact, fiction and forecast*, New York: Bobs-Merrill
- Gorard, S. (1998) The call for a middle-way in educational research, BERA Internet Conference, March 1998, [www.bera.ac.uk/debate/reply/g1.html](http://www.bera.ac.uk/debate/reply/g1.html)
- Gorard, S. (2000a) *Education and Social Justice*, Cardiff: University of Wales Press
- Gorard, S. (2000b) One of us cannot be wrong: the paradox of achievement gaps, *British Journal of Sociology of Education*, 21, 3, 391-400
- Gorard, S. (2001a) *The way forward for educational research?*, *Occasional Paper 42*, Cardiff: School of Social Sciences
- Gorard, S. (2001b) *Quantitative methods in educational research: the role of numbers made easy*, London: Continuum
- Gorard, S. (2001c) International comparisons of school effectiveness: a second component of the 'crisis account'?, *Comparative Education*, 37, 3, 279-296
- Gorard, S. (2001d) The ethics of robust reviewing, *Research Intelligence*, 76, p.14
- Gorard, S. (2001e) *The role of cause and effect in education as a social science*, *Occasional Paper 43*, Cardiff: School of Social Sciences
- Gorard, S. and Taylor, C. (2002) Market forces and standards in education: a preliminary consideration, *British Journal of Sociology of Education*, 23, 1
- Gorard, S. and Taylor, C. (2002) What is segregation? A comparison of indices in terms of strong and weak compositional invariance, *Sociology*
- Gorard, S., Rees, G. and Jephcote, M. (1998) The role of contour lines in school improvement, *Research Intelligence*, 66, 30-31
- Gorard, S., Salisbury, J. and Rees, G. (1999) Reappraising the apparent underachievement of boys at school, *Gender and Education*, 11, 4, 441-454
- Griffiths, M. (1998) *Educational research and social justice: getting off the fence*, Buckingham: Open University Press
- Hagenaars, J. (1990) *Categorical Longitudinal Data: Log-linear, panel, trend and cohort analysis*, London: Sage
- Hakim C. (1992) *Research Design: strategies and choices in the design of social research*, London: Routledge
- Hakuta, K. (2000) *Perspectives on the state of education research in the US*, presentation at AERA, New Orleans, April 2000
- Hammersley, M. (1997) Educational research and teaching: a response to David Hargreaves' TTA lecture, *British Educational Research Journal*, 23, 2, 141-162

- Hammersley, M. (2001) On Michael Bassey's concept of the fuzzy generalisation, *Oxford Review of Education*, 27, 2, 219-225
- Hammersley, M. (2001) *Some questions about Evidence-based Practice in Education*, presentation to BERA, Leeds, 13-15th September 2001
- Hargreaves, D. (1997) In Defence of Research for Evidence-based Teaching: A Rejoinder to Martyn Hammersley, *British Educational Research Journal*, 23, 4, 405-420
- Hargreaves, D. (1999) Revitalising educational research: lessons from the past and proposals for the future, *Cambridge Journal of Education*, 29, 2, 239-249
- Hay/McBer (2000) *Research into teacher effectiveness: Phase II report*, (mimeo via BERA office)
- Hayes, E. (1992) The impact of feminism on adult education publications: an analysis of British and American Journals, *International Journal of Lifelong Education*, 11, 2, 125-138
- Hendry, D. and Mizon, G. (1999) *The pervasiveness of Granger causality in econometrics*, Nuffield College Oxford, (mimeo)
- Hillage, J., Pearson, R., Anderson, A. and Tamkin, P. (1998) *Excellence on research in schools*, Sudbury: DfEE
- Hume, D. (1962) *On Human Nature and the Understanding*, New York: Collier
- Johnson, B. (2001) Towards a new classification of nonexperimental quantitative research, *Educational Researcher*, 30, 2, 3-14
- Kennedy, M. (1997) The Connection Between Research and Practice, *Educational Researcher*, October 1997, pp. 4-12
- Kuhn, T. (1970) *The structure of scientific revolutions*, Chicago: University of Chicago Press
- Lakatos, I. (1978) *The methodology of scientific research programmes*, Cambridge: Cambridge University Press
- Larvor, B. (1998) *Lakatos: An introduction*, London: Routledge
- Levacic, R. and Glatter, R. (2001) Really good ideas? Developing evidence-informed policy and practice in educational leadership and management, *Educational Management and Administration*, 29,1, 5-25
- Magee, B. (1972) *Popper*, London: Fontana
- Mahoney, J. (2000) Strategies of causal inference in small-N analysis, *Sociological Methods and Research*, 28, 4, 387-424
- Marshall, G. (2001) *Social Sciences*
- McIntyre, D. and McIntyre, A. (2000) *Capacity for research into teaching and learning*, Report to TLRP
- McKim, V. and Turner, S. (1997) *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences*, Indiana: University of Notre Dame Press
- Medical Research Council (2000) *A framework for development and evaluation of RCTs for complex interventions to improve health*, London: MRC.
- Millett, A. (1997) *Speech to TTA Research Conference*, 5 December 1997, London: TTA
- Moore, L. (1996) Statistical analysis for the evaluation of school health promotion. In Piette, D. (ed) *Towards an evaluation of the European Network of Health Promoting Schools*. Brussels: Commission of the European Communities & Universite Libre de Bruxelles.
- Moore, L., Paisley, C. and Dennehy, A. (2000) Are fruit tuck shops in primary schools effective in increasing pupils' fruit consumption? A randomised controlled trial, *Nutrition and Food Science* 30(1) 35-38

- Mortimore, P. (2000) Does educational research matter?, *British Educational Research Journal*, 26, 1, 5-24
- Mortimore, P. and Sammons, P. (1997) Endpiece: a welcome and a riposte to the critics, pp. 175-187 in White, J. and Barber, M. (Eds.) *Perspectives on school effectiveness and school improvement*, London: Institute of Education
- National Educational Research Policy and Priorities Board (1999) *Investing in learning: a policy statement with recommendations on research in education*, Washington DC: NERPP
- National Educational Research Policy and Priorities Board (2000) *Second policy statement with recommendations on research in education*, (draft for consultation)
- National Research Council (1999) *Improving student learning: a strategic plan for educational research and its utilization*, Washington DC: National Academy Press
- Pedhazur, E. (1982) *Multiple Regression in Behavioural Research*, London: Holt, Rhinehart and Winston
- Pirie, A. (2001) Evidence-based practice in education: the best medicine?, *British Journal of Educational Studies*, 49, 2, 124-136
- Popkewitz, T. (1984) *Paradigm and ideology in educational research*, London: Falmer
- Pötter, U. and Blossfeld, H. (2001) Causal inference from series of events, *European Sociological Review*, 17, 1, 21-32
- Prandy, K. and Bottero, W. (2000) Social reproduction and mobility in Britain and Ireland in the nineteenth and early twentieth centuries, *Sociology*, 34, 265-281
- Pring, R. (2000) Editorial conclusion: a philosophical perspective, *Oxford Review of Education*, 26, 3&4, 495-501
- Pring, R. (2000) *Philosophy of educational research*, London: Continuum
- Rees, G., Gorard, S., Fèvre, R. and Furlong, J. (2000) Participating in the Learning Society: history, place and biography, in Coffield, F. (Ed.) *Differing visions of a learning society*, Bristol: Policy Press
- Resnick, L. (2000) *Strengthening the capacity of the research system: a report of the National Academy of Education*, presentation at AERA, New Orleans, April 2000.
- Reynolds, D. (2000) *The pupil progress project: final report*, (mimeo via BERA office)
- Riddell, S. (1992) Gender and education: progressive and conservative forces in balance, p. 44 in Brown, S. and Riddell, S. (Eds.) *Class, race and gender in schools*, Glasgow: SCRE
- Roberts, I. (2000) Randomised trials or the test of time?: The story of human albumin administration, *Evaluation and Research in Education*, 14, 3&4, 231-236
- Roker, D. (1991) *Gaining the Edge: The education, training and employment of young people in private school*, London: City University
- Rom, W. (1992) Causation, in *Textbook of Environmental and Occupational Medicine*
- Salmon, W. (1998) *Causality and explanation*, New York: Oxford University Press
- Saunders, L. (200) Understanding schools' use of 'value-added' data: the psychology and sociology of numbers, *Research Papers in Education*, 15, 3, 241-258
- Scott, D. and Usher, R. (1999) *Researching education: data, methods and theory in educational enquiry*, London: Cassell
- Speller, V., Learmonth, A. and Harrison, D. (1997) The search for evidence of effective health promotion, *British Medical Journal*, 315, 361-363
- Taylor, E. (2001) From 1989 to 1999: A content analysis of all submissions, *Adult Education Quarterly*, 51, 4, 322-340

- Thouless, R. (1974) *Straight and crooked thinking*, London: Pan
- Tooley, J. and Darby, D. (1998) *Educational research: a critique*, London: OFSTED
- Travis, G. and Collins, H. (1991) New light on old boys: cognitive and institutional particularism in the peer-review system, *Science, Technology and Human Values*, 16, 3, 322-341
- TTA (2000) *The Teacher Research Grant Scheme: summaries from the second year of the scheme*, London: Teacher Training Agency
- Willinsky, J. (2001) The strategic education research program and the public value of research, *Educational Researcher*, 30, 1, 5-14
- Woodhead, C. (1998) *Academia gone to seed*, *New Statesman*, 26 March 1998, pp. 51-52