**'This is an electronic post-print version of an article published in *International Journal of Research and Method in Education* Vol. 30, No. 3 (2007): 307-323. *International Journal of Research and Method in Education* is available online at: http://www.informaworld.com/smpp/title~content=t713727792.**

URL to published version: http://dx.doi.org/10.1080/17437270701614790.

# Where does good evidence come from?

Stephen Gorard
School of Education
University of Birmingham
s.gorard@bham.ac.uk

Thomas Cook
Northwestern University

## Abstract

This paper started as a debate between the two authors. Both authors present a series of propositions about quality standards in education research. Cook's propositions, as might be expected, concern the importance of experimental trials for establishing the security of causal evidence, but they also include some important practical and acceptable alternatives such as regression discontinuity analysis. Gorard's propositions, again as might be expected, tend to place experimental trials within a larger mixed method sequence of research activities, treating them as important but without giving them primacy. The paper concludes with a synthesis of these ideas, summarising the many areas of agreement and clarifying the few areas of disagreement. The latter include what proportion of available research funds should be devoted to trials, how urgent the need for more trials is, and whether the call for more truly mixed methods work requires a major shift in the community.

## Introduction

This paper is unusual in that it has some of the characteristics of a debate between its two authors. The debate concerns what we consider policy-makers and practitioners require in the form of high quality education research evidence. Should education research be predominantly experimental or based on mixed-method studies? The main issue we address in this brief paper is the place of experimental design in evidence-based policy-

making and practice. One of us, Thomas Cook, has written widely on randomised control experiments in education (Cook & Campbell 1979, Cook 2002), detailing their merits, outlining the assumptions and threats to validity associated with them, and listing and refuting most objections usually raised to doing them. The other of us, Stephen Gorard, is probably better known for writing about the value of mixing methods of different kinds (Gorard 2001, Gorard with Taylor 2004). Our difference in research emphasis speaks to a vexing issue of method choice that currently bedevils the educational research community as it seeks to ground education policy decisions in better evidence. As will become clear, our positions are quite similar when we address general propositions such as the question forming the title to this paper. But we differ in some particulars of great importance for deciding which kinds of education research to commission in order to improve the policy yield of education research. Learning where we agree may help readers identify where they can be relatively confident about method choice. Learning where we disagree may help them identify which method choice decisions remain problematic and where maximal caution is required in evaluating claims about new knowledge for improving the outcomes from education. Each of us will first individually present some research principles and propositions, and we then later draw together our areas of agreement and difference. The emphasis throughout is on clarity of expression.

## Thomas Cook

*1. Educational policy speaks to many different kinds of issue and question, most associated with different method preferences. So, comprehensive "evidence-based research must be multi-method.*

Among other issues, educational policy has to be concerned with "who gets what?"; "what does a given educational service cost?"; "what is classroom life like?"; "how well are students performing?"; "how are teachers trained?"; and "what works to improve student performance?" The majority of these questions are descriptive; only the last is explicitly causal. Theorists of method in the social sciences broadly agree that the best methods for dealing with non-casual issues require theory, ethnography, interviews and surveys, among other methods. Experiments hardly help. If educational research is to speak to the comprehensive knowledge needs of the education policy community, it can, should and must involve multiple methods. Framing the issue as a choice between experimental or mixed methods is silly. Even questions that seem purely causal at first glance are embedded within contexts where we also need to know: "Who gets the new educational practice under evaluation?" "What does the program cost?" "Which social values does the intervention speak to?", and so on. Even the major institutional advocate of experiments today, the Institute for Educational Sciences of the United States Department of Education, routinely commissions experimental evaluations that also include theoretical analysis of the program under review and observational measures of program implementation. It also funds many, and some very large, non-experimental surveys of educational resources and performance, like the National Assessment of Educational Progress (NAEP). Arguing for mixed method research is anodyne, given the

heterogeneity of knowledge needs in education and the research design practices of even the most passionate advocates of experiments. The debate needs to be framed differently--about (1) the priority to give to causal versus non-causal issues in educational research today; and (2) when causal questions are central—and only then--the priority that should be given to randomised control experiments versus other causal methods. I address these two basic themes in the points below.

*2. Causal questions have a special importance in educational policy research.*

My rationale for this assertion is that policy-makers are selected or elected to make decisions. These decisions often touch on how to change schools and colleges to raise the performance of teachers and students. This is always a pressing concern, but especially in nations where comparative studies like PISA indicate disappointing levels of average performance. But even in nations currently doing well, novel ideas are needed if they are to maintain their relatively high standing. Where are these ideas to come from, and how should they be tested before being implemented on a broad scale? I believe many descriptive issues are important in education; but identifying "what works' deserves a special status among the concerns of those accountable for the quality of educational performance, as does learning about "what works" in the most secure ways. Moreover, I also believe that the need to learn what works is especially acute right now, raising even more the priority of gaining accurate causal knowledge in education. The main reason for believing this is immediately below.

*3. The causal knowledge now being generated in education is inadequate for providing a secure stock of knowledge about effective educational practice.*

Empirically based causal assertions are rampant in today's educational research, very few of them the product of experiments. How valid are they in general? No definitive answer is possible, given that an answer depends on the very standards of evidence that are in contention among educational researchers today. But in the countries I know best--the USA, UK, France and Germany--no secure body of literature exists that policy makers can rely upon to learn what should be changed in schools in order to improve student achievement and social behavior. Cacophonous claims about effective practices abound. But we will later see that their technical warrant is generally weak when evaluated by the most widely accepted causal methods in statistics and across the social sciences as a whole, as opposed to the standards currently operating in large parts of the educational research community. When the fundamental values buttressing policy choices are at issue, all educational policy-makers should welcome active dispute since contention about values is the mother's milk of democracy. But to welcome dispute about the effects of discrete educational practices is another matter. Evidence-based policy depends on a reasonably clear research-based consensus about effective practices as one central input into decision-making, though all decision-makers realize that total consensus is impossible. Yet typically decision-makers do not get even an approximation to consensus. Some decisions have been endorsed by education researchers in the past and were widely disseminated without much quality research evidence to back them up. Some of these turned out to be quite disastrous--e.g., new math and whole language

reading instruction. I believe that causal issues are central to educational policy and that the causal knowledge generated by educational researchers to date has generally not been trustworthy. So the key is to learn more about what works in education. One proposal to do this involves radically increasing the incidence of random assignment experiments, since in Cook's (2002) review of the relevant literature they constitute from 1% to 5% of all the educational research studies that claimed a causal finding. Why stress experiments?

*4. For answering causal questions, the randomised experiment is well warranted theoretically and empirically.*

The theoretical warrant for experiments comes from a minor variant on the same statistical theory that undergirds the highly successful survey research industry. This minor variant uses statistical theory to create, not a single sample that formally represents the population from which it was drawn, but two or more samples that represent the same population, whatever it might be. Since the groups so created are initially identical on expectations, any final difference between them must be due to whatever intervention one group had that the other (or others) did not. However, this is not the only warrant for experiments. Over the last decades we have had considerable experience implementing them in sectors other than education and even some experience in education, albeit primarily in the USA. We are fast learning how to improve their implementation in order to regularly meet all the assumptions the method requires. The survey research industry could not exist without both a statistical theory and decades of wisdom (much from small-scale experiments) about how to implement surveys so as to reduce bias. The needed statistical theory already exists for experiments, and knowledge is being quickly accumulated about how to implement them more often and better (Cook, 2002). I do not want to argue that experiments are perfect, only that they are superior to their current alternatives. Their imperfections are of several kinds.

*5. The valid causal interpretation of experiments depends on assumptions being met.*

To produce unbiased causal results experiments require several assumptions that are routinely described in method texts. The major ones are that a correct random assignment procedure is chosen; that it is correctly implemented; that no differential attrition occurs across the groups being compared; and that contamination of the intervention details from one group to another is minimal. Also, the analysis of experiments depends on standard statistical assumptions being met, as do other causal studies too. Each of these assumptions can be violated, but methodologists know about them and about how to avoid or limit their influence in complex settings like schools and colleges. However, while many educational researchers know about the necessary statistical theory, far fewer of them are experienced in implementing experiments so that their assumptions are demonstrably met in school-based research, and on a quasi-automatic basis (Cook & Foray, in press). Experiments are only sufficient for unbiased causal knowledge when the above assumptions are demonstrably met, and meeting them is not difficult for those with experience conducting experiments.

*6. Being limited in their capacity to generalize causal findings, experiments do not always answer the question of greatest policy relevance.*

Many experiments are limited to those schools, teachers or students that agree to whatever treatment they are assigned by chance (Cook, 1991). The causal findings so generated will be bias-free, but only apply to those who volunteer for a random assignment study. Other types of causal study will also depend on volunteers, but not necessarily volunteers of the same kind. Experiments have other restrictions to their generality. They do not guarantee that any obtained effects will hold in the future; and the effects of an intervention may change if it is implemented on a much broader scale that leads to different causal processes being involved in the smaller experiment than in the extrapolation to, say, an entire nation. Once again, though, these restrictions apply to varying degrees to other kinds of causal study too. The limited generalization of findings from single experiments helps explain why advocates of experiments prefer policy to depend on multiple experimental studies, each with a different population of persons, settings and times as well as on different ways of instantiating the intervention and measuring the outcome. Alternative causal methodologies are also limited in their capacity to generalize, although not all in the same ways as experiments. What are these alternatives? And how good are they? We must answer this to support the claim that experiments are marginally superior to their alternatives, albeit not perfect.

*7. In human history, valid causal knowledge has often come from non-experimental and non-quantitative sources.*

It would be preposterous to maintain that experiments are necessary for causal knowledge. Our ancestors learned about the causal effects of making fires millennia before there was formal experimentation. And scholars knew that out-group threats usually cause in-group cohesion long before R.A. Fisher created the first formalization of experimental design. The case for experiments is that they are needed for detecting effects that are smaller than many of the others humans have learned about in the past. We have learned from studies of educational performance net of various student background characteristics that, within the limits of the models used, schooling effects are indeed very small and swamped by individual differences, particularly familial and psychological ones, not to speak of the genetic ones still to be examined in detail (*Coleman, 1966; Jencks, 1972). This may be why the Institute for Educational Sciences designs its evaluations to detect achievement gains of $1/5^{th}$ of a standard deviation--typically over several years and thus equivalent to a total of about one year's change in growth over these years. As important as such effect sizes are, they are not obviously "large" and are manifestly far from transformational. Experiments are also needed because many educational practices that might be effective are enmeshed in real-world school or college life within complex systems involving many other variables. This makes it difficult to identify the unique causal role of any one educational practice, or set of practices, unless these practices have first been isolated and then systematically varied.

*8. In social science, experiments are not the only method known from theory to be capable of generating unbiased causal knowledge.*

Four alternatives to the experiment are known to generate unbiased casual inferences under certain conditions. (1) From statistical theory and comparative empirical research reviewed in Cook (2007), we know that regression-discontinuity studies can produce the same causal estimates as experiments. These studies depend on an educational resource being distributed according to an eligibility score along some quantitative continuum, often a specific level of need or merit but sometimes a specific date of birth or order of applying for the service under review. The key is that everyone on one side of the eligibility score receives the service and those on the other side do not. (2) We also know from theory that instrumental variable approaches can result in unbiased causal inference when an instrument is found that is correlated with the treatment but not with errors in the outcome (Angrist,  ). We also know that casual inferences are unbiased if (3) the process of assignment to treatment is perfectly known or (4) the outcome is perfectly predicted (Cronbach, 1982).

*9. These theoretically unbiased alternatives have assumptions that cannot be as clearly met in actual research practice, making them technically inferior to the experiment.*

Regression-discontinuity has less statistical power to detect effects than the experiment (Trochim, 1984), and it depends on strong assumptions about the functional form of the relationship between the assignment variable and outcome (Rubin, 1977). As for instrumental variables, it has proven very difficult to find many of them that meet the requirement of being uncorrelated with the outcome—the ironic exceptions being random assignment (Angrist, Imbrens & Rubin, 1996) and regression-discontinuity (Hahn, Todd & VanderKlauuwe, 2002). Most causal claims to date using such instruments, particularly in economics, have been hotly contested and thus limit our confidence that an instrumental variable approach can be widely used to promote causal inference. Both random assignment and regression-discontinuity derive their intellectual warrant from the fact that the process of assignment into the different treatment conditions is completely known and hence easily modeled. This is not the case with quasi-experiments or non-experiments where attempts are made to model the treatment assignment process. Empirical research on attempts to do this via selection models (Heckman, 1979) and propensity scores (Rosenbaum & Rubin, 1984) show that these statistical tools nearly always fail to recreate the results of experiments that share the same intervention group and so vary only in how their control group is formed—at random or not (LaLonde, 1984; Glazerman, Levy & Myers, 2002; Cook, Shadish & Wong, 2007). So full knowledge of the treatment assignment process has not yet turned out to be a viable and practical causal tool. And it is almost always impossible in actual research practice to totally predict any educational outcome, even when schools are the unit of study. The foregoing implies that the main case for preferring experiments is that they are practically superior to the other causal methods known from theory to be unbiased.

*10. Many other methods are also currently used for supporting claims about what works in education, but they are generally inferior because they do not enjoy an independent theoretical or empirical warrant as unbiased.*

A great array of other methods is used to justify causal claims in education. They range from site visits to countries that are performing well in PISA through to highly statistical difference-in-differences or causal modeling studies. Also included are ethnographic accounts, secondary analysis of survey data, and quasi-experiments. None of them enjoys an independent and theoretically infallible warrant sufficient to justify the causal knowledge gained. The shortfalls are many and vary by method. Suffice it to note here that Campbell and Stanley (1963) and its successors (Cook & Campbell, 1979; Shadish, Cook & Campbell, 2002) have detailed many threats to the validity of causal conclusions that are associated with even the better of these study types. Moreover, Glazerman et al. (2002) have documented how practice among economists, including some who work in education, regularly fails to produce the same results as experiments that share the same treatment group. The absence of both a theoretical and empirical warrant for the many types of study from which causal conclusions are regularly drawn in education today could well be a major reason why so many causal claims have failed to stand up to hard scrutiny and have not led to clear cumulative learning about what works.

*11. In many sectors where policy is currently made, experiments enjoy more credibility than other kinds of causal study.*

This is the case in health, public health, agriculture, the prevention sciences, criminal justice, and legal studies of compliance with gender- and race-based hiring laws. And even in survey research, improvements to practice have often depended on experiments. They are also common in research on early childhood education in the USA where Congress requested that its largest national program, Head Start, be evaluated experimentally. Also, the pre-school studies regularly cited to promote the "universal preschool" policy in the USA are held in such high regard because they are experimental and involve decade-long effects on children's lives (Weikart, Reynolds, Ramey, ). To advocate against randomised experiments requires a compelling argument that schools are systematically different from institutions in other sectors in ways that either make experimentation infeasible or bias the results obtained. Such advocates also have to explain why experiments are common both in pre-schools and in school-based research with prevention rather than academic achievement outcomes. It is important to note that experimentation does not exist in a vacuum.

*12. Any single experiment assumes prior knowledge that need not itself be the product of experiments.*

Experiments require prior substantive theory and the experience of persons knowledgeable about what is feasible in school life. They also require the availability of good measures of the preferred outcomes, or the ability to construct such measures. Further, they require at least local political and administrative support for the study. And finally, they depend on prior causal studies. These can be experiments, but need not be so in order to confer marginal advantages for constructing future experiments. For instance, statistical power calculations depend on variance estimates from other studies, as do bigger picture issues like how an intervention is conceptualized, chosen and implemented. All experiments build on the shoulders of prior scholars in theoretical and

applied fields. They do not exist in a methodological vacuum, and experimenters are not a new priesthood that can afford to declare itself independent of educational research' past.

*13. Having information from experiments does not guarantee that this information will be used in policy debates, and certainly not used to form a decision.*

Although experiments give a marginally superior causal answer compared to other methods, this does not guarantee that these results will be more often used in debates about educational change. And when evidence from experiments is used, it certainly does not mean that they will alone shape policy decisions. The history of educational research is replete with examples of study results not apparently used; and in democracies decision-making does, and should, depend on many factors other than scientific knowledge alone.

*14. But having scientific information from experiments probably increases the odds of the information being used in policy debates.*

It is difficult to argue this point for education today, given the recent history of school-based experiments with random assignment. However, in other fields of study, causal results from experiments are routinely preferred over the results from other kinds of study. This is especially true in medical, public health and prevention contexts, and also when the results from multiple studies are synthesized in search of an effective policy option. Indeed, it is standard practice in meta-analyses to analyze the results from experiments separately and to add non-experimental results to the review only if their average effect size does not differ from that from experiments (Cook, Cooper, ). This is even the case in those rare educational instances where a very large number of studies of an intervention exist, creating enough experiments to analyze separately even if they are but a tiny fraction of the whole corpus of studies— for two instance in early childhood reading, see Ehri (2001, a,b). In more qualitative review contexts, at least in the USA, expert panels commissioned to review the literature for a governmental agency often pay special attention to the experiments in formulating conclusions for policy consideration within a government agency, deliberately giving them more weight than the non-experimental evidence.

*In conclusion*, the argument is that learning "what works" is crucial in educational policy-making, and that it is especially a problem today. This is because we have failed over the last 30 years to accumulate a secure body of knowledge about effective educational practices. So I believe that the case for more causal research is clear--that is, relative to other kinds of study with a claim on educational research funds. To do more experiments does not mean that only experiments are valuable and that only they should be funded. But it does mean that they deserve, at least temporarily, a higher profile than they received over the last 30 years or so. But only if the causal studies provide more secure causal knowledge of what works, and the best method for achieving this involves doing experiments, given their independent warrant in statistical theory and also in past practice in sectors outside of school-based education. Experiments are not perfect. But no other

method currently exists that does as well, and this is broadly acknowledged in sectors other than education. But it is also acknowledged in two sectors with close links to traditional education—in research on cognitive outcomes in pre-schools and on prevention outcomes in research in schools. Experimentation is not a novelty in school-based research; merely something whose sphere of application needs to be extended to meet a commitment to learn more about what works in a context of international crisis about educational performance levels in many larger countries.

**Stephen Gorard**

Like Tom Cook, I shall set out a number of summary propositions. Interested readers can trace the further basis for these propositions in my research writings – examples of which are provided. In my own writing I am concerned with education as an area of public policy, including pre-school, post-compulsory, and adult, provision, whereas Tom Cook writes for the context of schools. I see no reason why this difference should affect our methods approach.

*1. A key ethical concern for those conducting or using publicly-funded education research ought to be the quality of the research, and so the robustness of the findings, and the security of the conclusions drawn.*

Until recently, very little of the writing on the ethics of education research has been concerned with quality. The concern has been largely for the participants in the research process, which is perfectly proper, but this emphasis may have blinded researchers to their responsibility to those not participating in the research process. The tax-payers and charity-givers who fund the research, and the general public who use the resulting education service, have the right to expect that the research is conducted in such a way that it is possible for the researcher to test and answer the questions asked. Generating secure findings for widespread use in public policy could involve a variety of factors including care and attention, sceptical consideration of plausible alternatives, independent replication, transparent prior criteria for success and failure, use of multiple complementary methods, and explicit testing of theoretical explanations through randomised controlled trials or similar experimental designs (Gorard 2002a).

*2. It is helpful to consider the research enterprise as a cycle of complementary phases and activities, because this illustrates how all methods can have an appropriate place in the full cycle of research.*

Experimental designs, like in-depth work or secondary analysis, have an appropriate place in the cycle of research from initial idea to development of the results. The main reason to emphasise experiments at this point in time is not because they are more important than other phases in the cycle, but because they represent a stage of work that is largely absent in education research. If nearly all of education research were currently conducted as laboratory experiments then I would be one of the commentators pleading for more and better in-depth work or secondary analysis, for example. Other weak points

in the cycle are currently the systematic synthesis of what we already know in an area of work, the design or engineering of what we already know into usable products for policy and practice, and the longer-term monitoring of the real-world utility of these products (Gorard with Taylor 2004, Gorard et al. 2004).

*3. Working towards an experimental design can be an important part of any research enterprise, even where an experiment is not envisaged or even possible.*

Sometimes a true experiment, such as a large randomised controlled trial, is not necessary, and sometimes it is not possible. An experiment is not necessary in a variety of research situations, including where the research question does not demand it, and where a proposed intervention presents no *prime facie* case for extended trialling. An experiment may also not be possible in a variety of research situations, including where the intervention has complete coverage, or has already been implemented for a long time, and where it would be impossible to allocate cases at random. However, a 'thought experiment' is always possible, in which the researcher considers no practical or ethical constraints except answering the research question as clearly as possible. In then having to compromise from this 'ideal' to conduct the actual research, the researcher may come to realise how much more they could be doing. There might then be more natural experimental designs, more practitioner experiments, and surely more studies with appropriate comparison groups rather than no explicit comparison at all (a situation which reviews show is the norm for UK academic research in education). There might also be more humility about the quality of the findings emanating from the compromise design (Gorard 2002b, 2003a).

*4. Part of the problem of research quality lies in traditional research methods training and 'experts'.*

In the UK, traditional methods training for new researchers in university departments of education generally starts by introducing students to differences between types of research, and emphasising the purportedly incommensurable values underlying the variety of approaches to discovery. Most obviously, researchers are introduced to a supposed paradigmatic division between 'qualitative' and 'quantitative' studies in a way that encourages methods identities based on a choice of only one of these 'paradigms'. This leads many of us to indulge in paradigmatic strife, or write off entire fields of endeavor – as being 'positivist', for example. Some commentators try to heal these schisms after they have been created, but there is a shortage of texts and training resources that take the far superior approach of assuming that there is a universal underlying logic to all research. Such an approach leads from the outset of training to a focus on the craft of research, thus bringing design, data collection, analysis, and warranting results to the fore, leaving little or no place for paradigms (Gorard 2003b, 2004a).

*5. Part of the problem of research quality lies in a lack of appropriate use of numbers.*

One of the main reasons why there is not more mixed methods education research is clearly that there are few researchers willing and able to work with numbers. Since experimental designs are seen by many, incorrectly, to be 'quantitative' in nature, this could also be part of the reason for the lack of experimental work. There may be a number of influences at play here, including poor maths teaching in schools, lower ability of social science students in comparison to other disciplines both in terms of maths and perhaps also overall, the selection of methods courses by students in terms of perceived ease, and the widespread misunderstanding that being a 'qualitative' researcher means never having to deal with numbers. However, I am coming increasingly to the view that a major share of the blame lies with 'quantitative' researchers. They seem to prefer devising more and more complex methods of analysis rather than devoting their energy to creating higher quality datasets that are easier to analyse. They often present their research in exclusive and unnecessarily technical ways. They generally assume, incorrectly, that numbering is the same as measuring, that reliability is the same as validity, that probabilistic statistics can be used with purposive samples or even with population figures, and that any use of numbers must be based on sampling theory. This is not the way forward (Gorard 2006a, 2006b).

*6. Part of the problem of research quality lies in an unwillingness to test our cherished theories.*

Another element of the methods crisis stems from our love of specific theories, and our consequent unwillingness to test them for failure. A typical piece of evaluation in UK education is either commissioned by, or conducted by, those responsible for the programme being evaluated. There may then be pressure from funders to 'finesse' the results. I have certainly been contacted by evaluators seeking some new kind of analysis that will gainsay the surface findings, and which will support instead their underlying belief that the programme must be being effective. This is no different, in principle, to the dredging of data that goes on shamelessly *post hoc* in other forms of research as well. I have also experienced far too many cases in which researchers simply make up or distort data in order to help preserve their prior beliefs. Some methods experts actually advise researchers to 'take sides' before conducting research, and not to publish negative or otherwise unhelpful results. Of course, it remains true that the evidence-based approach to policy-making and practice is itself untested in education, and still far from fully satisfactory in fields such as health sciences. But this is a reason to test it, not to reject it out of hand (Gorard 2004b, Gorard and Fitz 2006).

*7. Much of the solution lies in greater scepticism, because the problem is not really one of methods at all.*

Some of the criticism of education research during the 1990s was concerned with relevance. But education is a very applied field of research. I do not find much published research that has no relevance to some important or useful component of education. The criticism is more properly about the poor quality of much research, so that even though the findings may have relevance they still cannot be used safely. In response, capacity-building activities have tended to focus on solutions in terms of methods, such as having

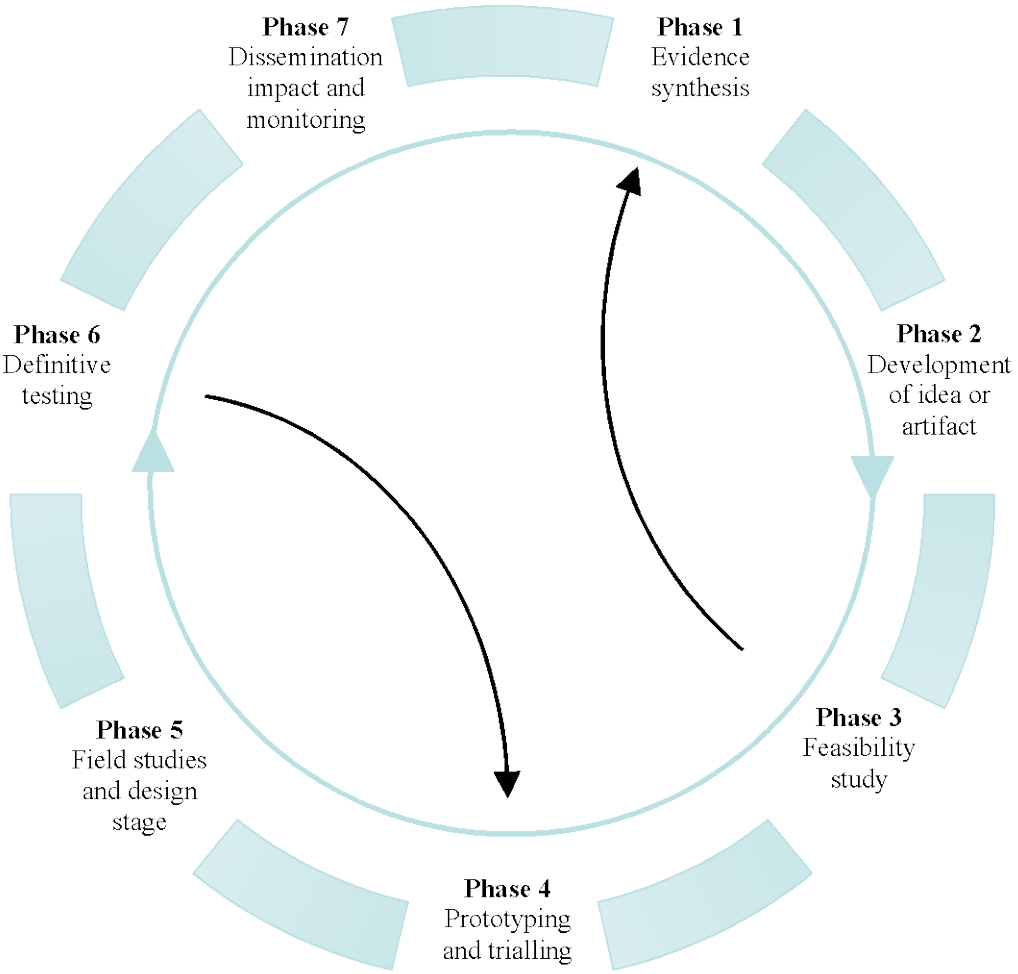
more complex quantitative work, more systematic reviews, or more experiments. These, to my mind, are not the answer in themselves. A more general change is needed in the culture of research. The answer for me lies in genuine curiosity, coupled with outright scepticism. These characteristics lead a researcher to suit methods to purpose, try different approaches, replicate and triangulate, and attempt to falsify their findings. It leads them to consider carefully the logic and hidden assumptions on the path from evidence to conclusions, automatically generating caveats and multiple plausible interpretations from the standard query – 'if my conclusions are actually incorrect, then how else could I explain what I have found?'. Some improvement may come from researcher development, but, somewhat pessimistically for an educator, I have come to believe that the role of capacity-building is limited here. Some people appear genuinely curious and sceptical anyway. Some, on the other hand, tend to be devoted 'believers' of things, and their development may involve simply a change of the subject of those beliefs as when a committed religious person becomes an enthusiastic Marxist, or when a 'qualitative' researcher turns heavily 'quantitative' (Gorard 2002c, 2005). In a sense, what we need for evidence-based policy making and practice is more real research, where the researcher is genuinely trying to find something out. From this, all else will likely follow – including more and better experiments for many of the reasons advanced by both authors in this paper so far.

## Agreements and Disagreements

Intriguingly, having written out our opening positions independently, it seems that we are mostly in agreement, though there are differences of emphasis we will mention. We agree that all commonly used methods have a valid purpose and a place in the larger cycle of education research. Our capacity-building should, therefore, focus on filling in the existing gaps within the cycle so as to create the needed expertise and practices, on trying to overcome mono-method identities where researchers reject the use of all but one type of evidence, and on teaching respect for all methods in their place, as difficult as it is to identify these places.

We also agree that the full research cycle represented in Figure 1 presents a simplified and stylized, but useful, model of the research cycle. In this cycle, reviews and secondary analyses might appear in Phase 1, theory-building and small-scale fieldwork in Phase 2, et cetera, with smaller experiments being part of Phase 5 and a full randomised controlled trial only appearing once, in Phase 6. We agree that experimental designs are not privileged for all of these phases and that other means are preferable, especially for the first four phases. We also agree that experiments are currently lacking in education research practice writ large, and that most education research gets stuck in phases 1 to 4. In other words, it is stuck working towards a randomised trial that hardly ever gets done.

Figure 1 – An outline of the full cycle of education research




We further agree that it is important to answer descriptive questions such as 'Who gets what?' or 'How are teachers trained?'. But these questions are no sooner broached than we usually also want to learn how to improve things in these domains and causal questions then arise, like: 'How can we train better teachers?' or 'How can we better share out resources? Thus, a complete programme of education research will generally lead to a need to make causal claims, and so to an ethical need for researchers to use something like a randomised controlled trial to make these claims responsibly.

Important consequences follow from our agreement that most education research gets stuck in phases 1 to 4 and that experiments have a special role to play in the underrepresented phases 5 through 7. For a fixed research budget, doing more experiments in the later phases will entail fewer resources for those researchers working on phases 1 through 4, this being the vast majority of education researchers. So these individuals will not, and do not, like increasing the priority accorded to causal questions and methods. This priority is deeply threatening to them intellectually and instrumentally, hence their lack of support for the call to conduct more school-based experiments

Drawing attention to the neglected later phases of the research cycle indirectly serves to raise the priority accorded to them. After all, there is little point to a model that rarely meets its ultimate goals! Without explicit or implicit priorities, Figure 1 is conservative in its implications. It is a recipe for more of the same since so few education researchers want to work on the later phases, or even know how to do so if experiments are required. They might want to argue that phases 1 through 4 are necessary for the later phases, thus justifying much more work on the earlier than the later phases, especially since the figure presumes a winnowing process - only some modest fraction of the ideas initially generated ever get to have a randomised experiment devoted to them later. However, we both agree that the early phases are not necessary conditions for the later ones, as advantageous as it is to have them. Indeed, many educational practices that are currently widespread have never been through even the first four phases of Figure 1. They are widely implemented despite theory that is weak or even non-existent and, if any studies support these practices at all, they are not strong in terms of internal or external validity, having mostly been conducted in contrived settings or tested in a few schools and with few classrooms or children. In the past, we have been accepting of educational reforms that have hardly benefited from phases 1 through 4, let alone 5 and 6. Even in logic, there is no need for potential school reforms to have gone through a multi-year testing process before being implemented in schools.

Also pushing towards conservatism is that an un-prioritized Figure 1 leaves the funders of education research with total freedom of action. They never need take stands about priorities, and so they need not fear alienating their constituencies in universities and ministries. In many policy environments, setting priorities is a political headache one would like to avoid if possible. Figure 1 may be a good normative description of some Platonic research cycle, but it will only change education research practice if it is linked to acknowledging two things we both agree on concerning its last phases—that they are: (1) indispensable to evidence-based policy research since much of policy is about improving educational performance; and (2) they are neglected in current education research practice, making secure knowledge about what works in education a current gap of some significance.

Where we may differ more is on the urgency of the need to fill this gap and hence on the extent to which experiments are needed. Tom Cook is more worried that current education research rarely gets to a point where it reliably tests its ideas in the hurly-burly of school life, and that so few organizations responsible for education and research on education are fazed by this. He believes that those commissioning education research have a responsibility for hurrying along the research cycle and for short-circuiting it on a regular basis by jumping quickly to stages 5 and 6. He argues that the last phases in Figure 1 are the *sine qua non* of evidence-based education research. Without them, policy-makers do not have secure causal evidence, arguably the most relevant of all kinds of evidence for forming policy. Consequently, policy-makers cannot truly meet their accountability obligation to tax-payers. Of course, there are always researchers willing to offer policy-makers causal knowledge; but without experiments they cannot offer causal knowledge that is known to be secure because it results from a valid statistical theory based on random assignment and from the wisdom about implementing experiments that

has accumulated from doing them in complex settings in the past, including even from randomised experiments on doing randomised experiments (e.g., Shadish, Clark & Luellen, 2007).

Stephen Gorard sees the need for more causal studies at the end of the research cycle in Figure 1, and also the need for more experiments in Phase 6. Indeed, he has supported both as Director of the ESRC-funded Research Capacity-building Network in the UK. This helped convince him of the difficulty of shifting the culture in UK higher education research, though he nevertheless continues to take on the task and is currently leading an ESRC-funded Researcher Development Initiative designed to promote the use and understanding of randomised controlled trials (http://trials-pp.co.uk/). However, he is less worried about the shortage of knowledge about effective educational practices than Tom Cook is; and he is also less sure of the size of the premium that experiments deserve when causal knowledge is needed. So he does not use the rhetoric of crisis and, if we were to re-assign some hypothetical education research budget, he might not assign as much money to experiments as Tom Cook would. However, this is a difference of degree rather than a fundamental difference about the relative importance of causal questions and experimental methods.

However, we do disagree on whether calling for more genuinely mixed methods is 'anodyne', as Tom Cook terms it. Stephen sees the dominance of qualitative studies in UK education journals and regrets the number of researchers who fail to accept the principle that different kinds of questions (phases) require different (multiple) approaches. Tom Cook sees different kinds of questions requiring different methods, but not each kind of question requiring multiple methods. For a given kind of question, one method is often superior to another. It is only across all of education research with its many different kinds of question that multiple methods are needed. And we both agree on this last proposition. However, Tom Cook sees it as so obvious that it is not worth claiming as an intellectual principle. In this sense, it is anodyne for him, however gripping the need for mixed methods may be as part of a political battle between research factions that struggle to be at the table for prestige, funds and self-vindication. But the main point is that we both agree that randomised controlled trials are the best available primary method for answering causal questions. We both want to know, therefore: How can we get more of them done and done well?

**References:**


Angrist, J.D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects using Instrumental Variables. *Journal of the American Statistical Association, 91*, 444-472.

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.

Coleman, J.S. (1966). *Equality of Educational Opportunity*.

Cook, T.D. (1991). Clarifying the warrant for generalized causal inferences in quasi-experimentation. In M.W. McLaughlin & D.C. Phillips (Eds.), *Evaluation and*

*education: At quarter-century* (pp. 115-144). Chicago: National Society for the Study of Education.

Cook, T.D. (2002). Randomised Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for Not Doing Them. *Educational Evaluation and Policy Analysis*, 24(3), 175-199.

Cook, T.D. (in press). "Waiting for Life to happen"; History of the Regression Discontinuity Design in Psychology, Statistics and Economics. *Journal of Econometrics.*

Cook, T.D., & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand-McNally.

Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Hedges, L.V., Light, R.J., Louis, T.A., & Mosteller, F. (Eds.). (1992). *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Cook, T.D., & Foray, D. (in press). Building the Capacity to Experiment in Schools: A Case Study of the Institute of Educational Sciences in the U.S. Department of Education. *Economics of Innovation and New Technology*.

Cook, T.D., Shadish, W.J., & Wong, V.C. (2007). When Non-experimental and Experimental Effect Size Estimates do and do not differ: A Review of the Within-Study Comparisons Literature. Institute for Policy Research, Northwestern University, Evanston, Ill.

Cronbach, L.J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Ehri, L., Nunes, S., Stahl, S., & Willows, D. (2001). Systematic phonics instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 3*, 393–447.

Ehri, L., Nunes, S., Willows, D., Schuster, B., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly*, *36*, 250–287.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus Experimental Estimates of Earnings Impacts. *The Annals of the American Academy, 589*, 63-93.

Gorard, S. (2001) *Quantitative Methods in Educational Research: The role of numbers made easy*, London: Continuum.

Gorard, S. (2002a) Ethics and equity: pursuing the perspective of non-participants, *Social Research Update*, 39, 1-4.

Gorard, S. (2002b) The role of causal models in education as a social science, *Evaluation and Research in Education*, 16, 1, 51-65.

Gorard, S. (2002c) Fostering scepticism: the importance of warranting claims, *Evaluation and Research in Education*, 16, 3, 136-149.

Gorard, S. (2003a) *Quantitative methods in social science: the role of numbers made easy*, London: Continuum.

Gorard, S. (2003b) Understanding probabilities and re-considering traditional research methods training, *Sociological Research Online*, 8,1, 12 pages.

Gorard, S. (2004a) Scepticism or clericalism? Theory as a barrier to combining methods, *Journal of Educational Enquiry*, 5, 1, 1-21.

Gorard, S. (2004b) Three abuses of 'theory': an engagement with Nash, *Journal of Educational Enquiry*, 5, 2, 19-29.
Gorard, S. (2005) Current contexts for research in educational leadership and management, *Educational Management Administration and Leadership*, 33, 2, 155-164.
Gorard, S. (2006a) *Using everyday numbers effectively in research*, London: Continuum.
Gorard, S. (2006b) Towards a judgement-based statistical analysis, *British Journal of Sociology of Education*, 27, 1, 67-80.
Gorard, S. and Fitz, J. (2006) What counts as evidence in the school choice debate?, *British Educational Research Journal*, 32, 6, 797-816.
Gorard, S., Rushforth, K. and Taylor, C. (2004) Is there a shortage of quantitative work in education research?, *Oxford Review of Education*, 30, 3, 371-395.
Gorard, S., with Taylor, C. (2004) *Combining methods in educational and social research*, London: Open University Press.
Hahn, J., Todd, P., & VanderKlaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*(1), 201-209.
Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-161.
Jencks, C. In F. Mosteller & D.P. Moynihan (Eds.), *On equality of educational opportunity*. New York: Random House.
LaLonde, R. (1986). Evaluating the Econometric Evaluations of Training with Experimental Data. *The American Economic Review, 76*(4), 604-620.
Rosenbaum, P., & Rubin, D.B. (1984). Reducing bias is observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.
Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics, 2*(1), 1-26.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.
Shadish, W.R., Luellen, J.K. & Clark, M.H. Propensity scores and quasi-experiments: A testimony to the practical side of Less Sechrest. In R. R.Boootzin (Ed.). *Measurement, Methods and Evaluation*. Washington, D.C.: American Psychological Association Press.
Trochim, W.M.K. (1984) *Research Design for Evaluation*. Beverly Hills, Ca.: Sage Publications.