Edward C. Page

What's methodology got to do with it? Public policy evaluations, observational analysis and RCTs

Book section

What's Methodology Got To Do With It? Public Policy Evaluations, Observational Analysis and RCTs.

Edward C Page

Are methodological choices critical to the success of an evaluation study? For policy evaluation research, the kind of research that governments and international organizations commission to find out whether policies or other interventions are working, one might expect methodology to play a more important role than in conventional academic research. If the questions evaluation research explores are relatively simple empirical rather than theoretical issues – above all whether the programme works or not, what is going wrong and how might it be fixed if not – and if governments make decisions committing huge public resources based on these evaluations, we might expect those who sponsor and conduct such research to be especially concerned with its scientific credibility as established through the empirical research techniques it uses (see Box 1). This certainly appears to be the reasoning behind those who advocate policy evaluation research adopting the "gold standard" of randomised controlled trials (RCTs), which are especially popular among politicians and government officials since they are deemed to be "the best way of testing whether a policy is working" (Cabinet Office 2012). Thus we may well conclude that methodology is important for policy evaluation.

Insert Box 1 Here (from end)

But the activity of evaluating policies is rarely simply a matter of developing and applying a convincing methodology to guide policy by showing government what works and what does not. This paper looks at the role of methodology in evaluations from the perspective of whether there is any evidence that policy makers are more likely to pay attention to, or act upon, studies that are deemed to be methodologically superior, whether by virtue of being more sophisticated, rigorous or appropriate. The concern of this chapter is not with establishing the merits and demerits of different methodologies in evaluation studies, but rather the role of methodology in explaining the impact or lack of impact of any evaluation studies. In practical terms it seeks to answer the question: if a researcher makes additional efforts to increase the integrity or sophistication of the research methodology used to perform an evaluation, will the effort pay off in terms of increased influence for that research?

This chapter first considers what a successful policy evaluation might look like and then goes on to consider the contribution that the level of methodological sophistication might make to that success. The generally small role that methodology plays as presented in these first parts of this chapter contrasts with the big role claimed by those advocating the adoption of RCTs. The fact that RCT methodology has been influential has more to do with its reputation for accuracy rather than any superiority of results that it produces. In the conclusion I go on to look at the problems of setting out a "gold standard" of evidence gathering for public policy evaluations and offer an account of the importance of methodology that reflects the wider constraints involved in evaluating public policy.

Success and the uptake of policy research

The common, if no longer entirely conventional, understanding of the success of policy evaluation research, here understood to be research commissioned by organizations with some view to shaping such policies (including terminating them), is related to its impact on policy makers and policy. In principle we can look at the impact of methodology in two stages: first on whether policy makers pay much attention to the research (uptake) and second whether this research actually improves the quality of policy or policy making. As will be seen, in practice the character of the first stage makes it difficult to assess the role of methodology in the second.

Policy uptake comes in three broad forms. First, a "linear" uptake where a specific piece of research having a discernable impact on a directly or indirectly related policy. This kind of uptake is extremely rare. While it is very hard to prove a negative, one can say that the most determined efforts to find evidence of specific pieces of evaluation research shaping directly the development of policy have long drawn a blank, irrespective of where and when it is sought. From the social research surrounding Great Society programmes in the US in the 1960s (Aaron 1978) to British local government "best value" and "evidence based" initiatives around the turn of the 21st century (Percy Smith 2002: 36) and UK national government in the early twenty-first century (Sowden and Raine 2008; Monaghan 2012) significant traces of a direct role of research in policy making have remained elusive. The literature on research utilization contains many convincing accounts of why research does not appear to be taken up by policy makers in this linear way (see Beyer and Trice 1982 for a meta-analysis), including those based on differences in timescales (Jowell 2003: 9-10), professional environments (Martin, Currie and Lockett 2011) and modes of argumentation between the worlds of science and politics (Ritter 2009) on and institutional constraints on policymaking (Waddell et al 1995), such that this lack of direct or "instrumental" influence can be described as overdetermined.

A range of scholars seeking to assess the impact of social scientific policy evaluations on policy tend instead to emphasise a second, less direct, uptake route: research can add to cumulative knowledge and understanding about the characteristics of policy interventions and can at some unspecified time be brought into the policy making process (Weiss 1977; 1995; for a review see Weible 2008). Research evidence does not have a direct "instrumental" use but can have less direct "conceptual" and "symbolic" uses in policy making (see Davies 2012). Carol Weiss' (1977: 534-5) definition of research as having an "enlightenment" function is, consequently, a widely accepted account of how social research affects debates about policy. It results from the "diffuse, undirected seepage of social research into the policy sphere" which "can gradually change the whole focus of debate" over a range of policy issues including education, housing, child abuse and legislative reapportionment (see also Weiss 1979).

A third uptake route, a political one, can be identified if one considers that some evaluations are commissioned for reasons that have less to do with providing an evidential basis for policy making than with politics. It is impossible to prove the *mens rea* issue of the intention behind commissioning evaluations, but one can point to evaluations that have served a range of political purposes including

- *endorsing the wisdom and foresight of the politicians who claim responsibility for a particular policy.* An evaluation of the "Troubled Families" (Casey 2012) programme, a form of "payment by results" scheme for local councils

making an impact on families with multiple social problems was claimed as an example "this government" turning "around the lives of thousands of troubled families".

- *advertising policies deemed to be successful.* An initiative to create "Sarah's Law", with the aim of informing parents when convicted sex offenders have access to their children was announced several times; including when a pilot was initiatied, when it was extended several times over and when it reported that it had "already protected more than 60 children from abuse during its pilot" (discussed in Goldacre 2010).
- *showing that politicians are prudent people who pay attention to evidence*. The British Home Secretary sought to use the small fig leaf of poor results from an evaluation to explain why she discontinued a widely ridiculed scheme to encourage illegal immigrants to "go home" by a campaign which included sending vans decked out with the "Go Home" message to advertise the scheme ("'Go home' billboard vans not a success, says Theresa May" The Guardian 22.10.13).
- *proselytising*. Research has played a significant role in helping persuade international organizations as well as other countries that a particular policy model should be adopted. Thus research evaluations helped make the case for international organizations supporting schemes such as the Directly Observed Treatment Shortcourse (DOTS) tuberculosis treatment programme and Conditional Cash Transfer programmes (see below).

Thus we have at least three broad mechanisms by which policy evaluations can be taken up, by: a) directly shaping the policy which it is evaluating (or one that is closely related to it); b) adding to the evidence illuminating how policies work and c) having an effect on perceptions of the policy makers or the policies they produce, an effect termed here "political". How important are methodological choices likely to be in each form of uptake?

Does methodology shape uptake?

What all three forms – linear, enlightenment and political – might appear to have in common is that they all rely on a significant degree of scientific credibility. This credibility might be bestowed on research by the application of conventional academic scientific rigour in developing and applying the methodology of the study. For the linear and enlightenment effects one has to have confidence that the results of the research are internally and externally valid to place any faith in them as a basis for discussing public policy, whether now or in the future. Moreover, using research for political cover or support will be less attractive if the methodology used to produce it is obviously full of holes.

The effect of methodology on linear impacts is hard to assess because of the sparsity of cases where such an impact is detectable. Education research has been one area that has generated sufficient studies for meta-analysis of research impacts. One of these (Cousins and Leithwood 1986: 346), which included impacts on instructors as well as policy makers, noted that most of the empirical analyses took potential for impact as a dependent variable rather than actual impact, moreover it pointed out that "increased methodological sophistication" appeared as likely to inhibit as increase the uptake of research.

We can point to an example of claimed linear impact where the UK Government stated that it "listened carefully" to, and acted on, to the findings from the National Evaluation of Sure Start, a programme with diverse components aimed at improving child welfare and education (DCSF 2008: 2). Did methodology play a role in determining which parts of the research were taken up? It is impossible to say as the document supposed to put such research based recommendations into effect is vague on the question of what particular research findings were listened to. The "Sure Start" guidance (DfE 2006) only ever mentions in rather general terms rather obvious points such as "Research has shown that many parents are unaware of the services on offer. It is important that centres make every effort to .. publicise the range of help they can give parents" (for a discussion of the limited impact of the Sure Start evaluation programme see Lloyd and Harrington 2012). Moreover it is not clear which of the recommendations to remedy such shortcomings are backed by research and which are not, and nowhere is the empirical basis discussed. Finding evidence of the linear uptake of policy research is hard enough, finding evidence that the uptake was at all affected by the methodological choices made in producing the research is harder still.

Moving on to the political uses of evaluations, one would not expect the influence of the methodological approaches used to produce the evidence on uptake to be strong, at least not above a basic level of credibility. Roberts, Petticrew, Liabo and Macintyre (2012) base their conclusion, "sound methods ≠ useable findings", on interviews with ten policy advisors from six countries with experience of handing evaluation evidence in dealing with politicians. One advisor argued "by and large, methodology is a weak influence in the sense that policy makers don't really tend to weigh up research evidence in terms of the strength of the source, it's much more the signal that they're interested in" and suggested that policy makers "tended to prefer very small scale studies, pilots, rather informal evaluation evidence where it supports what they're interested in doing, and [they are] quite resistant to the much stronger evidence where it doesn't support what they think.". Greenhalgh and Russell (2006: 36-7) endorse this view by suggesting that "social drama, personal testimony ("anecdotal evidence") is a uniquely authentic and powerful force" that can overrule hard statistical evidence.

Even in proselytising public policies cross nationally the quality of the research underpinning an intervention seems at best indirectly related to its uptake. For example, Walt shows how the research of Styblo, a Czech physician, in the 1970s was crucial in developing a form of treating tuberculosis and other diseases -- the Directly Observed Treatment Shortcourse (DOTS). But this research, important as it was in the community of technical health care specialists, did not reach the attention of World Health Organization policy makers for nearly two decades. The research was only taken up after a policy entrepreneur (an economist with experience and contacts in international health organisations) managed to package and sell it as a "broader, generalizable policy" (Walt, Lush and Ogden 2004: 199). Moreover, for the research on DOTS to be taken as a guide to international policy action required changes in the political environment, after "political elites in industralized nations became fearful that the disease would penetrate the ranks of their own middle classes, spurring the creation of a transnational coalition to fight the disease globally" (Shiffman, Beer and Wu 2002: 231). Methodology did play a significant part in this process, yet the policy environment played a far greater role.

Only in the illumination effect might one easily argue that the quality of the methodology matters, and this only by default. If the illumination effect is achieved in part by a piece of research standing the test of time -- being remembered and used in subsequent deliberations about desired policy options -- then it is at least a plausible hypothesis that scientific rigour will be related to the staying power of a piece of research. However, since there is no existing evidence to help us assess accurately how social research persists and shapes subsequent policy thinking or policy research, it must remain just a plausible hypothesis.

There is therefore overall little evidence to suggest that methodological choices affect the uptake of evaluation research. The notion that in linear models of impact (i.e. where a specific evaluation can be used to develop, modify or end a particular programme) uptake can be affected by methodological choices falls down in large part because evaluation research at best only rarely has such a direct impact. With more diffuse forms of uptake falling under the enlightenment model, the notion that methodological rigour will make the light from a good piece of research shine stronger and longer than that emanating from less impressive methodologies remains a plausible hypothesis, but one which carries little more weight than wishful thinking. In both linear and enlightenment forms of uptake it is hard to establish whether research produces better policy, let alone whether some methodologies produce better research which produces better policy, so we cannot really establish the effect of methodology of research on the quality of policy it helps produce. If we consider that evaluation research might have some form of political uptake, methodological rigour is one characteristic that can help establish its general credibility, but its political use is more likely to be shaped by a range of other features including the value of its findings to the politicians and others who seek to use it.

Improving uptake: the promise of RCTs

Given the generally low record of direct uptake of policy research findings, it is hardly surprising that many observers and policy makers have questioned whether there might be a better way of linking research to policy. It has become a widely shared view that better methodologies in evaluation research, specifically use of Randomised Controlled Trials (RCTs) would lead to better outcomes by providing harder and more accurate assessments of how well or badly a policy is working that policy makers would find harder to ignore. The basic idea of the randomised control trial is that one evaluates the impact of a policy intervention on the basis of comparing at least two groups to which those who are eligible to receive that intervention are randomly assigned. One group, the control group does not receive the intervention the other experimental group does. By comparing the outcomes for the control and experimental groups one can be confident of the precise impact of the intervention. This methodology is, of course, a conventional method of testing the efficacy of drugs and other medical interventions. The notion that the RCT is a higher form of evidence is most closely associated with "evidence based medicine" (see Shekelle, Woolf, Eccles and Grimshaw 1999). While the RCT has been used in social interventions for a long time (see Oakley 1998), it has become increasingly important as a method of evaluating government policies since the 1990s (Bassu 2013; Cabinet Office 2012).

The literature challenging the "gold standard" status of the method is now large and growing. Common criticisms include the expense of RCTs, recruiting and maintaining reasonably sized samples, the difficulties in avoiding contamination of treatment and control groups, the problems of external validity and the general criticisms that methodologies must be designed to fit research problems rather than specified independently of them (for a discussion of problems of RCTs in a medical context see Kaplan et al. 2011). Moreover evidence from medical trials suggests that observational studies do not necessarily produce results that differ from RCTs (see, for example, Benson and Hartz 2000). However, the concern in this paper is not with the methodological questions themselves but whether the method can buck the apparent trend of evaluation research not to be taken up. While there may be reasons for thinking that RCTs produce results that are harder to ignore and thus more likely to be taken up by policy makers, is there much evidence that this is the case?

The best evidence that this is the case comes largely through the proselytising mechanism of uptake. The discussion above on the spread of the DOTS tuberculosis programme it was an RCT (on treating sexually transmitted diseases in Tanzania) which helped foster international interest in the scheme even though it had earlier been highlighted through other forms of evaluation. The popularity of RCTs among policy makers seems to have added weight to the findings of other studies (though subsequent RCTs of the scheme have been less encouraging about its efficacy, see Tian, Lu, Bachmann and Song 2014). Another social intervention in which RCTs played an enormous part in proselytising was the development of Conditional Cash Transfers. Originally popularised through implementation of the PROGRESA programme in Mexico, the scheme linked welfare payments to conditions such as enrolling children in school. An RCT evaluation conducted by the International Food Policy Research Institute was especially influential in securing support from international organizations, especially the World Bank, in encouraging CCT schemes in other countries (see Handa and Davis 2006). In the international policy environment the methodological approach used by evaluation research can be an important component in selling an intervention.

At the domestic level, however, the promise of a linear uptake of RCT evaluations appears not to be fulfilled. It is hard to find examples of RCTs that have had direct impacts on policy. One of the biggest and most elaborate RCTs in recent UK experience was the Employment Retention and Advancement pilot evaluated over a seven-year period (see Greenberg and Morris 2005). It has produced significant volumes of government reports, working papers and published research on the substance of the programme, on issues relating to active labour market policy and on methodological issues in RCT evaluation. Yet a sympathetic appraisal of the impact of the evaluation on policy concluded that it had no "immediate or direct effect on welfare-to-work policies" but rather had "other effects in terms of informing and enlightening policymaking on welfare-to-work issues (i.e. a conceptual use of evidence)." (Davies 2013: R45). This finding, that well constructed RCTs on key policy issues do not affect directly policy development, is also echoed in work conducted on welfare-to-work policies in the US (Greenberg, Manell and Onstott 2000).

In part this lack of uptake might be because of the general tendency for RCTs to be more likely to show small or no effects for policy interventions, especially when one

looks for effects that last further beyond the immediate experience of the policy or programme. This tendency, noted in a range of studies of social experiments (reviewed for criminal justice interventions in Weisburd, Lum and Petrocino 2000) might be a result of Rossi's (1987: 4) "stainless steel law" of evaluation in general, that the "better designed the impact assessment of a social program, the more likely is the resulting estimate of net impact to be zero", or it might be a result specifically of the limitations of the RCT method (see Weisburd, Lum and Petrocino 2000). Either way it suggests that the prospects for uptake for RCTs might just as easily be lower than other forms of research evaluation as higher. This cannot, however, rule out uptake through negative mechanisms which are hard if not impossible to detect; such as when policymakers discard plans for a policy on reading an unfavourable evaluation.

Do we have any reason to think that policy based on RCTs is superior to policy based on observational studies? As uptake by policy makers in both is limited we do not have sufficient clear examples of either to be able to make a comparison. Perhaps the biggest intervention that has been supported by RCTs is the "cash transfer" scheme, in which the PROGRESA evaluation in Mexico discussed above played a large role, and which has become one of the most widespread development interventions of the past twenty years. Subsequent analyses of cash transfer schemes have certainly tended to find that their impact on a range of outcomes, above all in education, have been positive: that cash transfer schemes have an impact on the poor appears beyond doubt (see Baird, Ferreira, Özler and Woolcock 2013 although they also show there appears to be little difference between conditional and unconditional schemes). How far can it be considered that it was the evaluation of the PROGRESA scheme that generated good policy is, however, not so clear.

While the known defects of the PROGRESA evaluation are not of such a scale as to undermine its results, one can question how far it conformed to an RCT model. Faulkner (2014) describes how much of the implementation of the research was "murky". Many -- the evaluators, the Mexican government and international organizations among others -- had a political stake in establishing the story that this was an effective RCT. In the event it was best quasi-experimental, with evidence of non-random attrition rates as well as contamination across groups, the precise extent of which remains unknown. These problems aside, one can question what lessons policymakers outside Mexico took away from the study as what has been borrowed is not a single policy, but a range of cash transfer schemes, some conditional and some not, and where the conditions vary enormously. In short this looks less like the direct learning from a policy intervention through its RCT evaluation than the borrowing of a "label" or at best a broad idea (Mossberger 2000) of providing cash to very poor people, but for rather different purposes and implemented in a variety of significantly different ways and targeting different groups (see Sandberg 2015). This might be interpreted as policy makers themselves compensating for the oft-cited problem of RCTs, lower levels of external validity (i.e. the findings only apply under conditions identical to those where the study was conducted, not elsewhere) by devising their version of the policy based on the social, economic and political conditions prevailing in their jurisdiction (a point discussed in Pritchett and Sandefur 2013). We will never know how CCT schemes replicating exactly the Mexican model directly in different contexts might have worked, assuming such replication were even possible. Yet we can say that it is at least likely that the policies' success in other jurisdictions is due to

the non-RCT adaptation of the basic idea. The study helped spread what appeared and still appears to be a good idea, it did not itself provide much by way of guidance as to how it should be structured. One would need more than this to argue that the methodology itself produces good policy.

Methodology and evaluation

As Lindblom and Cohen (1979) argue, if we regard policy making as a problem-solving activity, the role for professional social science, above all that based on the application of empirical techniques of social inquiry, is likely to remain limited. Problem solving is generally based on "ordinary knowledge": basic understandings, ideas and beliefs about how the world works derived from "common sense, casual empiricism or thoughtful speculation and analysis. It is highly fallible, but we shall call it knowledge even if it is false" (Lindblom and Cohen 1979: 12). Not only this, but professional social inquiry, of the kind covered by empirical evaluation studies using social science techniques, are only made use of through the medium of such ordinary knowledge: ordinary knowledge helps define when professional social research is called for and how it is interpreted and used. To see the role of social science in policy problem solving as largely shaped by the methodology it uses, that better methodology, whether it be RCT or any other, is illusory. In Lindblom and Cohen's terms, this would reflect the quest to establish the "independent authoritativeness" of social research as a guide to problem solving. Social research might be, and often is, dependently authoritative because it supports or endorses ordinary knowledge. Emphasising the methodological sophistication of social science contributions to problem solving is seeking to give it a status and authority in the process that is independent of its relation to ordinary knowledge. All that we know about the utilization of research in policy making tends to underline Lindblom and Cohen's conclusion that this is an essentially vain quest.

The contribution of social science research to policy making does not have to be limited to the quest for authoritativeness through methodological sophistication. As Lindblom and Cohen (1979) go on to argue, there is a range of other contributions that professional social inquiry can make to policy-oriented problem solving including: conceptualising issues and shaping the intellectual frameworks of policy makers, providing evidence and argument, documenting what has been done in the past and with what result and challenging and changing ordinary knowledge. Insisting that social science's contribution should be largely a matter of applying only methodologies deemed to be of a higher order to weigh up whether a particular social intervention works or not is largely problematic in part because it hankers after an effect social science can never have and in part because it closes off the other possibilities for social science to make a contribution to problem solving.

None of this is to say that methodology is irrelevant in policy evaluation. By seeking to conform to high standards of professional social inquiry through adopting appropriate empirical research methods that can be justified and accepted by others in the same or related fields, social scientists do two things. First, they establish their credibility as people with something to say that could be worth listening to about policy issues. Second, they help establish their *locus standi* in the policy process. Their advocacy of different courses of action can be shaped, at least in part, by what they interpret to be the conclusions of their research. Some social scientists may not

need this *locus standi*: they may be advocates or zealots for particular policies, programmes or approaches before they go out into the field. For others their engagement with the field and their belief that scientifically valid conclusions deserve to be respected and the policy implications they draw from them be acted upon, gives them a basis on which they can become advocates. But even the greatest of methodologies will not guarantee that anyone will listen to them, and neither should it.

References

Baird S, Ferreira FHG , Özler B amd Woolcock, M (2013) "Relative Effectiveness of Conditional and Unconditional Cash Transfers for Schooling Outcomes in Developing Countries: A Systematic Review" Olso: Campbell Systematic Reviews 2013:8

Barton, S (2000) "Which clinical studies provide the best evidence? The best RCT still trumps the best observational study" *British Medical Journal* Jul 29, 2000; 321(7256): 255–256.

Basu K (2013) "The Method of Randomization and the Role of Reasoned Intuition The World Bank" New York: World Bank Policy Research Working Paper 6722

Benson K and Hartz AJ (2000) "A Comparison of Observational Studies and Randomized, Controlled Trials" *New England Journal of Medicine* 342:1878-1886.

Beyer, JM and Trice, HM (1982) The Utilization Process: A Conceptual Framework and Synthesis of Empirical Findings *Administrative Science Quarterly*, 27(4): 591-622.

Cabinet Office (2012) Test, learn Adapt. Developing Public Policy with Randomised Controlled Trials (London: Cabinet Office) (available https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf accessed March 20th 2015)

Casey, L (2012) "Listening to troubled families" London: Department for Communities and Local Government, July.

Davies, P (2012) "The state of evidence-based policy evaluation and its role in policy formation" *National Institute Economic Review* 219: R41-R52

DCSF (2008) "The Sure Start Journey. A Summary of Evidence" (London: Department for Children Schools and Families

DfE (2006) "Children's Centres Practice Guidance" (London: Deparment for Education and Skills) http://webarchive.nationalarchives.gov.uk/20100210152250/http://dcsf.gov.uk/everychildmatters/research/publications/surestartpublications/1854/

Faulkner, WN (2014) "A critical analysis of a randomized controlled trial evaluation in Mexico: Norm, mistake or exemplar?" *Evaluation* 20(2): 230–243

Goldacre, B (2010) “More than 60 children saved from abuse – small update” Bad science blog, August 7th http://www.badscience.net/2010/08/more-than-60-children-saved-from-abuse/#more-1748

Greenberg D, Mandell, M and Onstott, M (2000) "The dissemination and utilization of welfare-to-work experiments in state policymaking" *Journal of Policy Analysis and Management* 19(3): 367–382.

Greenberg D and Morris, S (2005) "Large-Scale Social Experimentation in Britain What Can and Cannot be Learnt from the Employment Retention and Advancement Demonstration?" *Evaluation* 11(2): 223-242

Greenhalgh T, Russell J. Reframing Evidence Synthesis As Rhetorical Action in the Policy Making Drama. *Healthcare Policy*. 2006;1(2):34-42.

Handa S and Davis B (2006) “The Experience of Conditional Cash Transfers in Latin America and the Caribbean” *Development Policy Review* 24(5) 513–536.

Jowell, R (2003) Trying It Out: The Role of ‘Pilots’ in Policy-Making, London: Government Chief Social Researcher’s Office, 2003

Kaplan BJ , Giesbrecht G, Shannon S and McLeod K (2011) "Evaluating treatments in health care: The instability of a one-legged stool" *BMC Medical Research Methodology* 11: 65 http://www.biomedcentral.com/1471-2288/11/65/ accessed 23.3.15

Lindblom CE and Cohen DK (1979) *Usable knowledge: Social science and social problem solving*. New Haven, CT: Yale University Press.

Lloyd, N and Harrington, L (2012) The challenges to effective outcome evaluation of a national, multi-agency initiative: The experience of Sure Start" *Evaluation* 18(1) 93–109

Martin, G Currie,G and Lockett, A (2011) "Prospects for knowledge exchange in health policy and management: institutional and epistemic boundaries" *Journal of Health Services Research & Policy* 16 (October):211—217

Monaghan, M (2012) "Cannabis Classification and Drug Policy Governance" in MacGregor S, McKeganey N, Monaghan M, Roberts M(eds) "Essays on the governance of drug policy" London: UK Drugs Policy Commission http://www.ukdpc.org.uk/publication/essays-on-the-governance-of-drug-policy/

Mossberger, K (2000) *The Politics of Ideas and the Spread of Enterprise Zones*. Washington: Georgetown University Press

Oakley, A (1998) " Public policy experimentation: Lessons from America" P*olicy Studies* 19(2): 93-114

Percy-Smith, J with Tom Burden, Alison Darlow, Lynne Dowson, Murray Hawtin and Stella Ladi (2002) "Promoting change through research: the impact of research on local government". (York: Joseph Rowntree Foundation). http://www.jrf.org.uk/publications/promoting-change-through-research-impact-research-local-government accessed December 2014

Pritchett L and Sandefur J (2013) "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix" Washington DC Center for Global Development Working Paper 336 http://www.cgdev.org/publication/context-matters-size-why-external-validity-claims-and-development-practice-dont-mix accessed 23.10.15

Ritter, A (2009) How do drug policy makers access research evidence? International J*ournal of Drug Policy* 20 (2009) 70–75

Roberts H, Petticrew M, Liabo K and Macintyre S (2012) "'The Anglo-Saxon disease': a pilot study of the barriers to and facilitators of the use of randomized controlled trials of social programmes in an international context *Journal of Epidemiology & Community Health* 66:1025-1029

Rossi P (1987) "The iron law of evaluation and other metallic rules". *Research in Social Problems and Public Policy* 4:3–20.

Sandberg J (2015) Transformative Social Policy in Development? Demystifying Conditional Cash Transfers in Latin America Lund University Expertgruppen för Biståndsanalys (EBA) Development Dissertation Brief 2015:03 (available eba.se/wp-content/uploads/2015/01/Johan-Sandberg.pdf accessed 23.10.15)

Shekelle PG Woolf, SH, Eccles M and Grimshaw (1999) "Developing clinical guidelines" *Western Journal of Medicine* 170(6): 348–351.

Shiffman J, Beer T, Wu Y.(2002) The *emergence of global disease control* priorities. *Health Policy Planning* 17(3): 225-34

Sowden SL, Raine R (2008) . Running along parallel lines: how political reality impedes the evaluation of public health interventions. A case study of exercise referral schemes in England. *Journal of Epidemiology & Community Health* 62 (9) 835 - 841.

Tian JH, Lu ZX, Bachmann MO, and Song FJ (2014) "Effectiveness of directly observed treatment of tuberculosis: a systematic review of controlled studies". *The International Journal of Tuberculosis and Lung Disease* 18(9):1092-8.

Waddell C, Lavis JN, Abelson J, Lomas J, Shepherd CA, Bird-Gayson T, Giacomini M and Offord DR (2005) "Research Use in Children's Mental Health Policy in Canada: Maintaining Vigilance Amid Ambiguity". *Social Science and Medicine.* 651:1649–57.

Walt G, Lush, L and Ogden, J (2004) “International Organizations in Transfer of Infectious Diseases: Iterative Loops of Adoption, Adaptation, and Marketing” *Governance* 17(2): 189–210.

Weisburd D, Lum CM and Petrosino A (2001) "Does research design affect study outcomes in criminal justice?" *The Annals of the American Academy of Political and Social Science* 578:50–70.

Weiss, CH (1977) "Research for Policy's Sake: The Enlightenment Function of Social Research Policy" *Analysis* 3(4): 531-545

Weiss, CH (1979) "The many meanings of research utilization". *Public Administration Review* 39, 426 - 431.

Weiss, C. H. (1995) "The Haphazard Connection: Social Science and Public Policy" *International Journal of Education Research* 23(2): 137–150.

BOX 1

Common basic designs used to evaluate policy

Randomised control trials (also known as "social experiments") in which some beneficiaries are randomly assigned to groups, one of which is a "control group" not exposed to the intervention and at least one other "experimental group" that is. The effect of the policy can in principle be assessed by comparing the control and experimental groups. These have become especially important since the 1990s and are often claimed to be the 'gold standard' of evaluation design (e.g. on the impact of different approaches to workfare on employment outcomes of those seeking work).

Before-and-after studies that seek to derive an understanding of the effects of a policy by inferring that they are reflected in changes over the *status quo ante*. A simple and effective design that can help assess the impact of a significant event or intervention (e.g. looking at the use of re-used shopping bags following a law mandating charges for single-use bags).

Area-based comparisons that introduce an intervention in some locations but not in others and assess the impact of the intervention by comparing the outcomes in the different places. Where systematically conducted these may approximate randomised control trials. Often also used to "pilot" programmes to see if interventions have any effect at all or whether there are problems in implementing them (e.g. proposed changes in unemployment benefit administration are tested in specific locations first).

Ethnographic studies that trace through impacts of policies by close observation of how those receiving and/or delivering the service behave. Useful among other things for exploring the reactions of poorly understood groups and unanticipated consequences (e.g. evaluating the impact of needle exchange programmes on the behaviour of injecting addicts).

Case studies that seek to trace through the impact of an intervention by following through a selection of cases (e.g. an evaluation of the value of evaluation studies based on tracing through the impact of a sample of such studies on policy).

Reputational studies that base their assessments of an intervention or policy on the perceptions of those receiving or delivering them. Not invariably to be dismissed as 'anecdotal' (e.g. where the policy seeks to change perceptions, such as in evaluations of programmes about the treatment of victims of crime).