Evaluation, evidence and review

Ian Greener

Introduction

It is frequently claimed that we live in a post-truth era where politicians play little attention to research evidence (Economist 2016). At the same time, however, they also appear to want to claim that their policies are pragmatic, common-sense views of ‘what works’ (Davies et al. 1999). Policy-makers often appear to want to cite anecdotes rather than evidence as supporting their reorganizations, but even the term ‘evidence’ conceals a great deal, so it is important to unpack what our expectations are around evaluation and evidence so that we are clear about what we are aiming for.

One starting point is to claim that evidence and evaluation should be linked to a scientific approach to policy-making. The problem is working out exactly what this means. Economists often advocate cost-benefit methods as a scientific approach, and it has become increasingly common to hear demands for a greater use of experimental methods (Mulgan 2015). At another extreme, evaluations can often appear to be an afterthought, and driven by politics rather than rigour.

This chapter considers the questions of what counts as evidence in evaluation, exploring three approaches that attempt to provide a method for the synthesis of research, with each either attempting to provide an underpinning for the way in which either interventions can be evaluated, or existing research synthesised for policy purposes (or both). It considers the strengths and weaknesses of each, as well as the assumptions they hold about what counts as good research. It concludes by suggesting that the realistic approach put forward by Ray Pawson has the greatest potential for informing policy through research because of its more sophisticated consideration of the problems of applying social research to policy ends, but that its application and widespread use will require a new mindset from both policy-makers and academics that recognises the limitations of what in research can be achieved, and which gives a greater role for creativity in social research.

Randomised controlled trials, evaluation and evidence

Many researchers regard the randomised controlled trial (RCT) and systematic review as representing a gold standard for the creation, assessment and evaluation of evidence. It is not hard to
see why. RCTs originate in medicine, presenting a paradigmatic case of the application of scientific methods. The idealised view of medicine combines rigorous research with informed professionals, and its achievements, as even its critiques accept (Le Fanu 1999), are truly astounding.

However, the extent to which what we might call the RCT model of evidence collection and assessment evaluation outside of medical textbooks is more of an open question than is often assumed (Greenhalgh 1998). In order to meet the requirements of RCTs, scientists undertake large-scale trials making use of double-blinded groups, with participants of the research selected on a random basis. These three elements (random sampling, large-scale trials, double-blinding) are attempts to capture the effects of treatments or other interventions in the most rigorous manner possible, and mean that results aim to be as generalisable as possible in relation to their population. However, the extent to which significant amounts of medical practice conform to these evidence standards is hugely variable.¹

When several RTCs have been conducted in the same area, provided that their methods are seen to be robust enough, they can then be carefully combined to form a meta-analysis to effectively merge the trials into a single, large-scale synthesis. This requires a great deal of care, with existing trials being closely scrutinised with only those trials meeting the highest standards (in terms of their reported compliance with the large-scale, randomised, double-blinded approach) being incorporated in full into the review. Clear inclusion and exclusion criteria are drawn up to aid those conducting the evaluation in deciding which studies should and should not be incorporated, and, if they are to be included, on what basis, and their effect sizes carefully examined.

The assumptions underlying the RCT

The RCT approach carries with it a range of ontological and epistemological assumptions that need to be considered carefully to assess whether they can be translated into social research. In terms of ontology, the methods treat the existence of an objective world as, perhaps a little paradoxically, both straightforward and problematic.

The use of double-blinded methods means that both researchers and participants are acknowledged as potentially having an impact upon research outcomes should they become aware of whether they are participating in the control or intervention group, and so implicitly acknowledging that mental constructs about the assumed efficacy of treatments or clinical interventions can influence the success of the research.

At the same time as acknowledging subjectivity in relation to treatment through double-blinding, however, the RCT approach treats the measurement of clinical outcome relatively unproblematically. The extent to which the efficacy of the intervention being assessed by research will undoubtedly be affected by a range of factors outside the control of scientists, as human beings go about their everyday lives doing a range of things that can improve or reduce the efficacy of the treatments under consideration, but make use of a relatively simplistic constant conjunction model of cause and effect.

The RCT method attempts to get around the problems of confronting open systems, and of variations in both intervention and subjects through the robustness of its method, by applying large-scale random sampling to try to reduce confounding factors, and by using its double-blinded methods in an attempt to reduce subjectivity. This suggests a view of evidence that is context-independent, but which is achieved by attempting to put in place controls over the experiment itself. This approach is generally also shared by social science disciplines, such as economics, which present themselves as scientific. The key question is the extent to which these
methods and ideas can be applied in social research more generally, or are specific to areas such as medicine where scientific methods may be applied (albeit imperfectly).

Problems with applying RCT methods in social science

The application of context independence, closed systems, and constant conjunction cause and effect measured by statistical methods, all have significant problems when transferred without adaptation into social science.

Context independence raises profound questions for social research. This is because many contemporary methodologists (e.g. Flyvbjerg 2001) suggest that the key to functioning in social systems is the acquisition of knowledge that is context-sensitive, or even context-specific. Some everyday examples illustrate this – even for professionals doing ostensibly the same job, moving to another place of work will mean learning new rules, new ways of doing things and a range of new social practices in order to be effective.

Context specificity occurs because (to bring in the second point above) we do not operate and live in closed systems where experimental controls can be applied, but the impact of our environments upon our lives is profound. This creates something of a paradox in biomedicine because whereas RCTs incorporate double-blinding to attempt to avoid subjective influences on interventions and treatments, medicine then has to confront subjectivity in its practice. For the vast majority of treatments, the patient’s social subjectivity and social context will have a profound impact on their clinical outcome.

The third (and again related) point is that, in social systems, measuring cause and effect statistically does not take any account of reflexivity (Giddens 1993), or of the possibility of more complex relationships between cause and effect. If I tell a colleague they are doing a good job today, it may have an effect on raising their moral. If I tell my colleague they are doing a good job tomorrow as well, the effect may be reduced but still occur. If I tell them every day for the next week, the effect may be diminished to almost nothing. If I continue to do it long enough, the effect may be close to zero or even negative. We are not machines or microbes, and respond in different ways at different times to different people, and repeating the same intervention is likely to produce a different result on different occasions (even where the intervention is identical), specifically because of the previous interventions. It is therefore extraordinarily difficult to evaluate the success of interventions in organisations that are primarily interpersonal.

A constant conjunction model of causation may also fail to capture relationships where causes and effects do not occur in a straightforward, repetitive manner. In the social world, cause and effect may take years to work through and be contingent upon a range of factors, but with causes still having important effects. Social class, for example, does not determine someone’s social mobility, but will lead to either the addition or subtraction of barriers in a person’s life in a way that is causally significant, even if it plays out differently for different people (Sayer 2005).

RCT research tends to presume that interventions lead to outcomes in relatively straightforward ways, where patterns between the causes and effects can be accurately and straightforwardly measured. But a new policy intervention will often affect different groups in different ways, and at different speeds and times. Increasing unemployment benefits, for example, may have complex effects on different groups of people, on the labour market more generally, as well as on incentives to work or remain on benefits, with effects taking months or even years to work through, and with different impacts in different regions of the economy (see also Chapter 32).

If, in social research, it is not possible to treat context as independent, to approximate a closed system in order to isolate the effects of an intervention, or to measure the result of interventions unproblematically, then it is not clear how it is possible to claim to have demonstrated whether
something ‘works’ or not through the application of methods derived straightforwardly from science.

**Natural experiments**

In countries such as the United States or Australia, different states may adopt different approaches to policy, making comparison between those interventions possible, and so some policy writers suggest they form ‘natural experiments’. There is some truth in this, but even then, the different interventions in different states occur in different social contexts (it would be hard to compare the results of different interventions in two US states, no matter how closely matched, without also taking account of their different histories, ethnic, gender and age make-ups, as well as other differences that exist across a huge range of other variable factors). In the UK is has become fashionable to argue that the increasing devolution between England, Scotland and Wales creates opportunities for cross-boundary learning, as each can be regarded as participating in a kind of experiment, but this is to ignore the very different historical, socio-economic and cultural contexts in each of those countries.

As such, despite their importance in biomedical research, the RCT-based approach has considerable problems when applied to the evaluation of social research, and to the construction of evidence-based policy advice. What alternatives are there?

**Adapting RCT approaches for social research**

A second approach to the problem of evaluating and synthesising social research comes in work that attempts to set quality standards through which studies may be excluded or included from evidence syntheses, but which accept that social research is often carried out in a different way, and with different aims, to RCT research.

There are several examples of work in this vein, especially from qualitative health researchers (Dixon-Woods et al. 2008). These approaches, while being inspired by RCT research, attempt to adapt both their criteria for inclusion as well as their synthesis method to take account of methodological differences. This leads to an attempt to find some means of applying statistical techniques to explore quantitative research and then using qualitative research findings to ‘triangulate’ with the quantitative synthesis to give an evaluation of the available research as a whole. One approach for attempting to achieve the goal of combining qualitative and quantitative research for synthesis has been pioneered in the work of Harden and Thomas and the research teams they work with, and has led to them creating what they call a ‘mixed method’ approach for ‘combining diverse study types’ (Harden and Thomas 2005).

Harden and Thomas present what they call the ‘standard stages of a systematic review’ (p. 259), in which ‘user-driven’ questions and boundaries are set for the study, a review protocol developed, a comprehensive search carried out, inclusion criteria applied and research quality assessed, after which data are extracted from papers for inclusion in the review, and findings synthesised, according to complex, usually statistical procedures. They aim to produce a qualitative (rather than quantitative) synthesis by reviewing qualitative work through similar stages as for quantitative research, going through the application of inclusion criteria, quality assessment, data extraction and then producing a synthesis of the qualitative findings.

Whereas quantitative assessments tend to use statistical techniques, the qualitative synthesis approach is based on the development of analytic themes, which can then be compared to the quantitative synthesis to seek matches, ‘mismatches and gaps’ (p. 264) towards a complex end result that blurs the distinctions between the two types of research in the synthesis, and instead
leads to (according to the authors) a means by which more diverse research questions can be answered to gain a better understanding of the phenomena. The researchers adopt what they call a ‘half-way house’ (p. 268) between the two different forms of data.

In asking a diverse range of questions, some of which may be answered through quantitative data, and some through qualitative, a more rounded synthesis can be produced that can attempt to find the best that both types of research can produce. However, significant problems will remain with this approach.

First, Harden and Thomas’ approach to synthesis tends towards the view that there is one right answer to the phenomena under investigation. This is, in many respects, something of a relic of the authors’ own research interests, which are mostly about assessing the efficacy of interventions that combine the social and the clinical. By assuming that the product of a research synthesis can produce a single right answer, there is an implicit assumption that interventions can be assessed in an acontextual way. However, as we have already seen, in social research this may be a significant problem – it may well be that context is all in such a situation. It is difficult to imagine a policy intervention that will not be significantly contextually affected in implementation. Equally, there is a real need to acknowledge the conflicting and complex nature of social life, and so that policies will be interpreted, understood and implemented in completely different ways (Stone 2012) – assuming that there is one right answer seems extremely hopeful.

Second, by putting in place selection criteria for the inclusion of qualitative research, the review method depends on there being consensus about what counts as good research, and that methods are clearly and consistently reported in social research publications – and neither is the case. Especially in policy research, relatively little space is often given in reporting research methods, and because there is less of a tradition of using such research for synthesis, work is not reported with a view to being included in syntheses later. Equally, there is no single ‘best’ methodological paradigm in social research, and so it is extremely difficult to draw up criteria against which we can review methodological validity – instead we have to assess work in terms of the research tradition it attempts to position itself.

Finally, it is not clear what would happen in Harden and Thomas’ model if the quantitative findings produced answers that were directly at odds with the qualitative research. Because the assumptions underlying their synthesis approach appear to treat the social world as existing independently of our concepts of it in a relatively unproblematic way – that we can straightforwardly measure the results of interventions in social contexts and work out which is the best – it is not clear what would happen if the results of their two different syntheses ended up contradicting one another.

In all, then, Harden and Thomas’ work looks promising in trying to break down the barriers between quantitative and qualitative research to produce policy-relevant findings, but instead ends up treating qualitative research as being synthesisable through a similar means as quantitative work, being quality assessable and reviewable using the single method they call ‘thematic review’. This seems to go against the grain of the standard justification for carrying out qualitative research (to ‘get close’ or produce contextually sensitive understandings of the social world), and so to end up treating it in terms different to those who carried out the work in the first place would have envisaged.

Realistic evaluation and review

A third approach to both evaluation and review comes from the work of Ray Pawson and his collaborators (Pawson 2006; Pawson et al. 2005; Pawson and Tilley 2001). Pawson has applied his ideas to a range of research problems, first coming to international attention in relation to his
work on evaluation (Pawson and Tilley 2001), but since moving on to a range of other insights,
especially in relation to considering evidence in policy research (Pawson 2002a, 2002b). Paw-
son’s work presents an extended critique of conventional approaches to evaluation and sys-
tematic review which share many of the same concerns offered above (and from which the above
work draws) – and from which he concludes that in attempting to adapt a method derived from
the scientific, experimental paradigm to social systems there is a risk of making a number of
methodological errors that carry the risk of misunderstanding the phenomena under investiga-
tion (Pawson 2013).

Rather than attempting to apply RCT principles to social research, Pawson suggests that it is
necessary, when considering social systems, to explore the complex interrelationship between
contexts, mechanisms and outcomes. In order to achieve this, a comparative approach needs to
be taken which looks at the way the same or similar mechanisms are introduced in differing
contexts to different outcome effects. This is a different causation logic to experimental methods
(which have in place a constant conjunctive model, where cause and effect are tightly coupled
and statistically measured), but instead through a generative model where causes and effects are
far more difficult to ascertain, and require a careful and creative unpicking of data in order to
unearth knowledge.

One of Pawson’s classic examples of exploring causality in this more complex way comes in
his explorations of ‘naming-and-shaming’ policies, where different contexts produce entirely
different outcomes (Pawson and Tilley 1997). Naming and shaming is based on the idea that
publishing the names of individuals who are deemed to be falling short of the behaviours gener-
ally accepted in society will lead to communities censuring those individuals, as well as prevent-
ing individuals from wanting their offences publicised. This is what we might call the ‘programme
theory’ for the policy. In the case of the use of Anti-Social Behavioural Orders (ASBOs), for
example, attempting to ‘name and shame’ offenders led to ASBOs not being seen as a potential
source of community censure, but instead as a kind of badge of honour, with some people even
competing to see who could get an ASBO at the youngest age, or who could receive the most
ASBO notices. In the case of paedophile registers, however, naming and shaming led to an
entirely different outcome.

What this points to is that policies based on the same programme theory can have very different
results because different social contexts will shape their effects in very different ways. Instead of
assuming that interventions are acontextual and that one policy intervention will work in all cir-
cumstances, the approach leads to more nuanced understanding of what works for whom in what
circumstances, and to the idea that we must explore the outcome effects of different contexts to
evaluate a programme. Evaluations need to take careful account of the complex relations between
outcomes and contexts, finding creative ways of exploring how the same programme theories
have been implemented through similar mechanisms in different contexts.

Pawson’s approach attempts to deal directly with the problems and concerns of the RCT-
based approach to social research explored above – that RCT-based approaches to evaluation
and review attempt to produce acontextual knowledge when social systems are almost inher-
ently contextual in terms of the knowledge required to navigate and understand them, that
whereas RCT-type work tends to try and approximate closed systems through the use of con-
trols, such an approach is not possible in social systems, and that causality in social systems needs
a different model to the statistically measured constant conjunctive model. In Pawson’s approach
it is necessary to be focused on trying to achieve a comparative, pattern-seeking approach that
looks for similar interventions across a range of cases to try to find out what they have in
common and differences in terms of both context and outcome, rather than trying to find sta-
tistical regularities among cases of a similar type.
Pawson’s approach, however, is not without its problems. One issue is that it can be extremely difficult to define what is meant by ‘context’ within the approach. This is because it appears to require the researcher or reviewer to be able to understand the range of occurrences within which the particular phenomena under investigation (the mechanisms) occurs, and then to work back and find differences between those occurrences in order to work out a system for differentiating them. This requires, in turn, that we are able to define a programme theory which is suitably abstracted in order to find a range of cases to compare (as in the case of naming and shaming above), but not so general that we end up comparing mechanisms that don’t have very much in common.

To take a very topical example, policies which attempt to ‘nudge’ people towards better outcomes (Thaler and Sunstein 2008) are extremely difficult as they cover such a wide range of potential interventions designed to engineer decision-making to encourage individuals to choose the ‘right’ outcome for themselves. Nudge interventions range from putting fruit next to tills in supermarkets rather than chocolate to try to encourage healthier snacking, through to requiring smokers to re-register annually with their doctor in order to have a licence that allows them to purchase cigarettes, as suggested by the former UK government health policy adviser Julian Le Grand.

Evaluating ‘nudge’ represents a potentially fascinating research project, but would have to include a very wide range of interventions that may appear to have little in common, and so require a creative process of comparison. It is striking that the House of Lords’ attempt to evaluate nudge programmes (House of Lords Science and Technology Select Committee 2011) failed to really get to grips with these challenges, and so reached the conclusion that ‘non-regulatory measures used in isolation, including “nudges”, are less likely to be effective’ (p. 5). However, this is because the nudge programmes were evaluated in a largely atheoretical and acontextual way – a more theoretically driven approach that too much greater account of context might have come to a different answer, and provided us with a much richer understanding of nudge interventions as a result.

There is also potential in combining realistic methods with those of the costs and benefits of societal interventions, based around a ‘social investment’ approach (Millar 2012). Such approaches engage stakeholders in planned changes, and so are likely to be multi-perspectival and to generate different understandings and theories of the social phenomena where change is being attempted. In monetising its findings, the social investment approach is controversial in social policy, but that has not prevented advocates from incorporating it into government when given the opportunity (Sunstein 2016).

Taking a contextual, programme-driven view of what is relevant in terms of context allows us to be careful in defining programme theories so that they are not so abstracted to be meaningless, and yet not so concrete that they do not allow comparative cases to be considered. It allows us to assess and compare the patterns of context, mechanism and outcome across a range of related activities to understand better how they illuminate the phenomena under investigation. In short, in Pawson’s approach, context, programme theory (mechanism) and outcome are necessarily interrelated, and it can be difficult to break into these interrelationships at an appropriate level to conduct the evaluation or review.

Pawson’s approach requires the researcher or reviewer to be able to find programme theories that have been applied in different contexts, and to be able to find the relevant contextual elements that appear to cause differences in outcomes. This will not be the same in each programme, and often requires trial and error before the best ‘fit’ is found (an example of a ‘retroductive’ research strategy). This may mean that different reviewers or evaluators have the potential to produce different evaluations. This is entirely understandable when we consider...
that the research is taking place in complex social systems whose workings may be open to
different interpretations, but can be alarming for policy-makers and research funders who may
prefer the apparently more concrete answers to which RCT-based research can lead. The key
question, however, is whether such research is oversimplifying the social because of its (relat-
ively uncritical) use of experimental scientific understandings about the world in contexts where
they may not apply.

Conclusion

Looking across the three approaches to evaluation and review explored above, all have their
problems. The RCT approach makes assumptions about the world that may not apply in social
research, as they assume a ‘one-best-way’ approach to methods and inquiry, along with a relativ-
ely closed environment and constant conjunction model that simply don’t reflect the reality
of conducting doing research in other settings. It is certainly difficult to argue from a philo-
sophical base that experimental methods hold much traction in complex social environments
(Byrne 2003) – and, even where experiments can be adapted for use in social research, they need
to be handled with extreme care to make sure their assumptions can stand up to the very
different research context rather than being assumed to be the only right way to do work.

The approach exemplified by Harden and Thomas is pragmatic in suggesting that RTC-type
methods can be adapted to become more multidisciplinary and incorporate qualitative research
findings, but in doing so they appear to attempt to assume a standardisation of the reporting of
methods in qualitative research that is unsustainable. They end up treating qualitative data in a
standardised way that seems to assume standards in the reporting of methods and which ignores
the standard philosophical justification of their use in achieving legitimacy through their close-
ness to research contexts, rather than through claiming to produce objective results and gener-
alisable findings. The end result is a standardised process that may appear to be multidisciplinary,
but has perhaps made rather too many concessions to the experimental RCT approach without
taking sufficient account of research contexts or the complex, open systems within which social
research often has to be conducted.

Pawson’s approach is both the most intellectually coherent in terms of its account of the
social world, but is also potentially the most variable in use in that it requires a great deal of
imagination and care in its use to find programme theories at the appropriate level to test, and
to be able to isolate the appropriate contextual factors making the difference to outcomes in
those theories across a range of cases. The need for creativity is likely to lead to differences in
interpretation among researchers, and appears odd to those researchers grounded in the stand-
ardised methods adopted by the other two approaches explored here. Pawson’s approach
requires us to place more faith in the researchers conducting the review or evaluation, but also
for them to earn trust by being clear and transparent about the method choices they
have made.

This chapter has argued that, in order for social research engaged in evaluation and review for
policy purposes to have the best chance of being intellectually coherent, its method must take
account of the different demands that social research has in terms of the complexity of the phe-
nomena under investigation, the degree of openness of the system under investigation, and the
type of causality which social systems have in place. Adapting methods which assume linear, con-
stant conjunction causation, based on assuming a relatively closed system in which interactions are
relatively simple seems to be misplaced for most social research. Attempts to adapt methods that
start from these assumptions, as Harden and Thomas do, seem destined to fall short, as they do not
start with a range of assumptions that match the worlds we are investigating.
Of the three approaches explored here, Pawson’s has the best chance of succeeding, but requires us to think of the role of the researcher or reviewer rather differently. Instead of asking them to be neutral, objective, independent observers who gain their legitimacy from an almost technical application of a particular method to their data, Pawson’s approach requires us to understand that the complexities of social systems are open to different interpretations that require creativity to explore, and intellectual honesty to explain, which understandings were chosen as the best and why. We have to put our trust more in the researcher and reviewer, and accept that any interpretation they come up with will be open to contestation – but that should actually be a part of the process of deciding which policies our politicians follow, rather than being merely a response to what they have already decided. Pawson’s work is both a response and a challenge to conventional approaches to evaluation and review, and should be treated as such, but in turn be open to challenge from researchers coming up with different interpretations of the evidence.

Following Pawson’s ideas requires us to accept that it is unlikely, in most social contexts, that there is a one, best answer. Instead there are multiple possibilities that are constrained by resources, ideology, social inheritances and a range of other factors. Policy debates should not be about pretending in such complex environments that we can find technical fixes, but instead exploring the interrelationships which new and existing research appears to highlight, and to engage both communities and policy-makers with our findings to try to make things better.

It seems that we can either find new ways of thinking about evaluation, accepting its limitations and that there is necessarily an element of creativity and subjectivity in their interpretation, or we can simply conclude that it is impossible to evaluate anything. Accepting the latter course of action is to give way to a passive form of postmodernism in which it becomes impossible to know anything for sure. We can accept the limits of our knowledge, but at the same time present findings not as definitive answers, but as our best attempts to understand the complex phenomena we are investigating, and as a basis for the beginning of dialogue.

Evaluation needs to deal with the messy, complex open systems into which policies are introduced. In such circumstances there are no final answers, but that does not mean that we shouldn’t be trying to put forward, honestly and openly, the best answers we can, accepting that they will always be theory-laden, but being transparent about our assumptions rather than attempting to conceal them behind a cloak of objectivity which is ontologically and epistemologically unsustainable.

Note
1 http://clinicalevidence.bmj.com/x/set/static/cms/efficacy-categorisations.html provides an outline of this.

References