Energy Efficiency (2018) 11:997–1008
https://doi.org/10.1007/s12053-018-9618-8

ORIGINAL ARTICLE

Evaluating energy efficiency policy: understanding the 'energy policy epistemology' may explain the lack of demand for randomised controlled trials

Adam C. G. Cooper

Received: 16 May 2016 / Accepted: 14 January 2018 / Published online: 26 January 2018

Abstract Vine et al.'s (2014) call for more randomised controlled trials (RCTs) in government-funded energy efficiency policy evaluation practice raises timely questions about what constitutes effective designs for evaluating and informing energy policy. Their implicit hypothesis that policy organisations share the same epistemic perspective as they do, and that the reason there are few RCTs are due to a set of barriers to be overcome is examined in relation to the UK government Department of Energy and Climate Change. Drawing on the author's experience of working in the ministry, the claim that barriers are a reason for preventing RCT use is discounted. An alternative explanation is presented, framed around the idea of an 'energy policy epistemology' that legitimately places certain specific knowledge demands and ways of knowing on research and evaluation designs. Through examination of a specific set of research and evaluation outputs related to the UK energy efficiency policy called the 'Green Deal', aspects of the proposed 'energy policy epistemology' are elucidated to explain the lack of demand for RCT designs. Final consideration is given to what kinds of designs are more likely to gain support in this context that might also deliver many of the benefits attributed to RCTs with longitudinal panels being one important example.

A. C. G. Cooper (✉)
UCL Department of Science, Technology, Engineering and Public Policy, 36-38 Fitzroy Square, London W1T 6EY, UK
e-mail: adam.cooper@ucl.ac.uk

Keywords Experimental design · Evaluation · External validity · Policy · Energy efficiency · Randomised controlled trials · Epistemology

Introduction

In their recent paper, Vine et al. (2014) make the case for promoting greater use of experimental research designs such as randomised controlled trials (RCTs) in the evaluation of energy efficiency policy. Their principal position is to 'advocate for the *primacy* of RCTs in testing the efficacy and/or effectiveness of various types of interventions by energy efficiency program administrators' (p. 629). Much of their focus is on public policy administrators, whom they identify as offering resistance to the use of RCTs and therefore need to be mandated to deploy them, for the betterment of society. This implies that the resistance in such settings is in some sense irrational that requires intervention to fix (not unlike the model taken by policy to address the uptake of energy efficiency technologies). Here, an alternative perspective is offered—that of the academically informed ex-policy research administrator proposing alternative, rational reasons why RCTs are not commonly used in energy policy. This opens up the possibility that a particular 'energy policy epistemology' is in operation that needs to be understood before failings within that framework can be identified. This results in an arguably better (or more potent) position to advocate in order to better enable more of the benefits of good research to inform policy-making.

Vine et al. are not the first to call for using this class of research design in evaluation policy practice (e.g. Haynes et al. (2012)), though it is one of the few calls for such an approach to be applied more widely within energy policy specifically (see Frederiks et al. (2016) for another). Within the UK government, 2010 heralded the arrival of the Cabinet Office's Behavioural Insights Team and with it a push to deploy more RCTs in policy analysis (Halpern 2015). This team affected the practice of policy evaluation within the UK government then Department of Energy and Climate Change (DECC) while the author served in DECC as the Head of Social Science Engagement (2011–2013). Their intervention gave rise to at least three RCTs or quasi-experimental designs during that time and provoked some debate within the department regarding the use of RCTs in policy evaluation. The deployment of RCTs and related experimental designs in policy evaluation has proved controversial in policy areas beyond energy—the adoption of RCTs within UK criminal justice policy-making in the late 1990s led to a small revolution within the UK evaluation community embodied by Pawson and Tilley's publication of *Realistic Evaluation* which was as much a manifesto for theory-driven evaluation as it was a rail against the application of RCTs (Pawson and Tilley 1997). Despite recent attempts to bring realist evaluation approaches together with RCT and experimental designs (Bonell et al. 2012), the controversy continues (Marchal et al. 2013) and the compatibility of RCTs with policy-making should not be taken as a given (Ettelt et al. 2015; Ettelt and Mays 2015).

Vine et al.'s diagnosis—that the energy efficiency policy sector lacks sufficient deployment of experimental designs for evaluation—has both heritage and controversy. Yet their argument is sufficiently persuasive to provoke closer inspection. Is energy efficiency a case of a sector which could benefit much more from RCTs than it currently does? The position presented here—drawing largely on the direct experience of the author as a policy researcher and evaluator combined with the observable behaviour of a national policy institution—is that their analysis misses important considerations which result in advocating RCT at the expense of pushing for other arguably more effective approaches, to the same end. Philosophically, the argument developed here is that to promote RCTs for energy policy evaluation is to force a square research design peg into a round 'policy epistemic' hole. Further, there are better ('rounder') approaches that deal with the thrust of Vine et al.'s points, but also help ensure better data for energy policy is made more generally available supporting innovation to address energy and climate challenges.

Below, Vine et al.'s case is briefly presented and their barriers to the deployment of RCTs reviewed with reference to the author's experience and observations of his time in the UK government.

The case for more RCTs in energy policy evaluation

Vine et al.'s case rests on a largely 'in principle' approach to understanding knowledge generation for policy. Logically, the argument runs, if policy is to have access to the best evidence, the best evidence must be the evidence subject to the least challenges to validity. Validity-challenged evidence is evidence that cannot be relied on to enable the right conclusions about what to do to address particular problems. Since, in principle, the use of randomisation in experimental designs provides the strongest defence against major threats to validity (alongside other safeguards such as the use of placebos, blinding the allocations and so on), the randomised control trial is the safest way of generating the best evidence for policy. Much of the power of this argument comes from an implicit idea that since RCTs are used in medical research, and since medical research has clearly provided many benefits to humanity, the use of RCTs must therefore by extension be good in other domains. However, even in the medical research, the use of RCTs is far from either uncontested or seen as a 'gold standard', as Bothwell et al. (2016) elegantly demonstrate. In medical research, RCTs are seen as just one kind of evidence—mainly in pharmaceutical settings—and other evidence (from qualitative case studies to observational epidemiology) are more commonly utilised in practice. This reflects exactly the observable research and evaluation practice of UK government departments. The question posed here then is as follows: are the lack of RCTs in energy policy due to 'barriers to deployment' as Vine et al. argue, or is the observable pattern of (non-RCT-based) research designs a reasonable (if imperfect) response to epistemic demands of policy-making? If the latter, what are those epistemic demands, and what can be done to improve evidence-making for energy policy? Is there any place for the forms of research design underpinning RCTs in policy research, and if so, where? The first question to be addressed is whether the barriers identified really are

barriers. One test of the barrier hypothesis is to observe whether RCTs emerge in situations where the barriers are absent.

Vine et al.'s barriers

Vine et al. identify four domains where barriers to the deployment or RCTs and related experimental designs exist. They call these barrier domains *regulatory*, *institutional*, *design* and *scope/theory*. Their Table 2 (p. 634) provides a useful summary of what they mean. Setting aside fundamental issues of whether a policy can be subject to an RCT (part of what Vine et al. call 'Scope/Theory' barriers), their Table 2 identifies five distinct 'barrier types', which are summarised as in Table 1.

This is not the first attempt at setting out barriers to accessing good evidence for policy-making. For instance, Oliver et al. (2014) undertook a systematic review of the research on barriers and facilitators to the use of evidence and identified the top five barriers and top five facilitators. In their study, Oliver et al. identify the five most frequently mentioned barriers. These overlap with Vine et al. around costs and skills. They do not mention risk aversion or equity/equal access issues. Both of these issues are particular to the deployment of resources for a new study, rather than access to data or information from an already completed one (the main focus of Oliver et al.'s study). Likewise, Oliver et al. mention *timing* as a key barrier, which Vine et al. do not. This latter one is important—a recent survey[1] of academics, policy officials and practitioners from the public and third sector carried out by UK-based charity National Endowment for Science, Technology and the Arts (NESTA)'s Innovation for Growth Lab found four major reasons for not using RCTs: lack of resources and/or time (23%), knowledge of RCTs (18%), equity/fairness (14%) and risk aversion (in the sense of concern about negative results, so similar to Vine et al.'s meaning, 11%). This again is a large overlap with Vine et al.'s barriers, though the small numbers reporting risk aversion and equity raise questions about how general these barriers are. The only major difference is the mention of *time*. Timing of research is likely a critical issue in policy-making, a concept that goes back to Kingdon's seminal work on 'windows of opportunity' (Kingdon 1984) and which is raised as a continual issue in evidence for policy-making circles (e.g. Bowen and Zwi (2005)). Timing issues are explored further below.

Table 1 Vine et al.'s barriers to deploying an RCT

Barrier type
1. Equal access to treatments/anticipated response to withheld treatment
2. Costs of running an RCT
3. Skills, capability and experience of staff responsible for demanding RCTs
4. Risk aversion to testing 'long established relationships and understandings'
5. Spillover or indirect effects are important

For each of these barriers, Vine et al. offer solutions which should help address them. This in itself begs the question: if the barriers are so obvious, and the solutions so straightforward, why are not RCTs a more regular feature of the design landscape in public policy commissioned evaluations? Is the major barrier of time a more potent one? Are there other reasons why RCTs are not a central part of a policy organisation's evaluation design considerations? These ideas are tested with reference to an indicative case study analysis of the evaluation approach taken by the UK government's Department of Energy and Climate Change (DECC) to a major energy efficiency policy (called 'the Green Deal') between 2011 and 2015 in the context of direct departmental engagement by the UK's Behavioural Insights Team.

Testing Vine et al.'s barriers hypothesis: the case of DECC's Green Deal Evaluation

In the UK, there have been only 3 recent published examples of RCTs in energy policy (DECC 2013, 2014a, 2014b[2]) directly paid for by an energy ministry. This contrasts with the (for instance) 17 separate (non-RCT) empirical studies on just two RCT-able UK energy policy programmes: the renewable heat incentive and the Green Deal/Energy Company Obligation (ECO) evaluations which were published around the same time.

[1] The survey was online, self-selecting (n = 170) and comprised responses from academics (21%), public sector (35%) and third/charity sector (37%). They reported the findings online in September 2016 here: https://www.nesta.org.uk/blog/barriers-experimentation-survey-results. Accessed 30 October 2017

[2] These are found via a search for 'randomised', 'trials' and 'controlled' in the www.gov.uk website, limiting the returns to Department of Energy and Climate Change. Accessed 30 October 2017.

This is in line with Vine et al.'s general observation regarding the US context. If Vine et al.'s barriers are both necessary and sufficient conditions to prevent the use of RCTs in DECC, then the author should be able to report conditions from DECC that align with the existence of the barriers, and thus explain the lack of RCT use. An absence of these barriers implies that there are other reasons for the relative lack of RCTs in this setting.

Did DECC lack the skills or experience to deploy RCTs?

The absence of a systematic survey of skills in the department precludes any strong conclusions here, but there are indications that DECC had sufficient skills available to deploy RCTs, at least in domestic sector energy policy. Firstly, as someone trained in experimental psychology and with a role to engage social science in DECC, the author was available to help design such a study should help be needed—but no requests were forthcoming. The internal cadre of around 15–20 social researchers, mainly focused on the consumer/domestic end of the energy system, were likewise available, though not all had backgrounds in experimental design. In addition, 90 or so economists (a social science discipline that particularly promotes RCTs) were evenly distributed to every policy team around the department. Further, the presence of the BIT in and around DECC meant that there was easy access to external advisors should such help be required. For the Green Deal team in particular, this flagship policy had both a dedicated evaluation specialist working in the team, access to the central Customer Insight Team in DECC, and worked with the BIT on a loft clearance trial (DECC 2013). The implication of this circumstantial evidence is that DECC did have (and continues to have) access to RCT designers on the Green Deal (and more broadly), but despite that, few RCTs are being or have been commissioned.

Did DECC lack funding to deploy RCTs?

The question of costs is implicitly a question of priorities. Clearly, the department had (and continues to have) significant resources to run even a large-scale RCTs. For the Green Deal and ECO policy in particular, a dedicated webpage[3] on gov.uk provides a neat summary of the work commissioned for this policy. This page shows 10 different substantive areas of developmental research and evaluation (excluding 'working papers'), any one of which could have been an RCT on its own. This implies that if costs played a role in preventing the use of RCTs to evaluate the impact of the Green Deal, it was also because it was such low priority that the design was at least ranked 11 out of a list of 11 possible approaches.

Did equity reasons prevent RCTs use?

The fact that DECC actually sponsored some RCTs with the help of the Behavioural Insights Team implies that equity considerations in these circumstances were not in general a consideration. It is possible that such a reason might be put forward with regard to the Energy Company Obligation element of the Green Deal and ECO policy since running an RCT on this aspect of the policy would inevitably mean that a set of low-income householders (the target of ECO) who would have received a free insulation package from their energy provider would not have received it (or at least it would have been delayed) in order to generate the control group. This reason could have been presented had the proposal to run an RCT been put forward. But the easy remedy to this is to focus on the more contentious area of policy—the main Green Deal, which was targeted at higher-income householders and was voluntary in nature. Yet no RCT was forthcoming.

Were officials risk-averse or afraid of negative results?

There are grounds for seeing this as a possible important reason. The wider context of the Green Deal is that the impact assessment for it suggested a large falling-off of installations of insulation and other measures with the introduction of the policy. This analysis was suppressed within the final published impact assessment, as projected installation measures were graphed only from 2013 onwards, following the introduction of the policy and therefore disguising the projected drop-off (DECC 2012, p. 10). However, at the same time, a range of studies were undertaken in support of Green Deal ex ante evaluation as outlined above. Surveys of customers, analysis of the Green Deal assessment and so on were all capable of returning 'negative results'. Of course, the argument might run that an RCT would have been much more potent in its assessment of the impact of the Green Deal and thus more problematic to handle

[3] https://www.gov.uk/government/collections/green-deal-and-eco-evaluation. Accessed 30 October 2017

politically. This argument remains an untested possibility.

Was an interest in spillovers important?

This is a difficult concept to assess. It is not completely clear what Vine et al. mean by 'spillover' as a barrier per se. In the experience of the author, UK policy officials working in energy had a general interest in understanding what is termed 'wider impacts', 'side effects' and so on. The economists in particular have an interest in these as evidence of such wider impacts (positive or negative) are potentially important considerations for the official impact assessment of the policy. In theory, as Vine et al. point out, there should be no a priori reason that an interest in spillovers should subsequently lead to a direct downgrading of interest in RCTs. What they do point out, however, is that the capturing of wider impacts via an RCT 'will require a very careful experimental design' (p. 637). This need to be 'very careful' implies a direct increase in complexity of commissioning of any such study which may then create a barrier to RCT deployment on account of the need to produce timely findings. Time and timing are now looked at as an additional important consideration.

Timing as a barrier

Policy-making can happen very quickly. The Green Deal policy that is being used as a central case example to illustrate key points here was launched in September 2013 after a development period of around 2 years. In 2015, just 2 years after launch, the Green Deal was closed down. Such pace demands high-quality evidence ready at the point of use if it is to stand any chance of influencing policy decisions. A study that takes a year to conceive, commission, set up, collect data and analyse would lose out to a study that can be turned around in 6 months. This counts against the deployment of more complex RCTs alone (as outlined above), since the level of co-ordination required between policy teams and analysts internally, the complex commissioning of such a study, negotiations around what is deliverable in the time and budget available, the complexity of analysis and availability of the best research suppliers are all problematic in this scenario. The response to this—to commission very simple RCTs—is also not without risk. Of the three RCTs carried out in 2013–2014 with DECC, two reported no effects (instead detailing the problems in delivering the interventions, DECC (2014a, 2013)) and one reported a significant effect but could not explain it (DECC 2014b). These findings in themselves show that even if an RCT can be delivered, it may not necessarily deliver useful outputs if its sole purpose is to capture an effect size between a treatment and control group.

This brief case study of DECC and the Green Deal in particular suggests that the barriers Vine et al. proposed do not in and of themselves provide a strong explanation for the lack of interest in deploying RCTs by energy policy institutions—at least in the UK. This implies that there are alternative drivers of evaluation designs in such settings, leading to specific patterns of commissioning. Below, a first attempt is made to set out what these drivers might constitute, drawing on the particular context of policy-making, and thereby provide an alternative possible explanation for the low incidence of government-commissioned RCTs in energy policy. Below, a closer examination of the kinds of designs that are prioritised by policy institutions is used to shed light on an alternative explanation.

Understanding the 'energy policy epistemology'

If energy policy-making makes different epistemic demands on research, these should be visible by examining the pattern of designs deployed to inform a policy. For clarity, these demands will be referred to from hereon as the 'energy policy epistemology' as the arguments made here are meant largely to focus on the particular demands of evaluating energy policy. The one proviso of such a deductive approach is that any such pattern may be subject to the risk aversion to negative findings identified by Vine et al. and others, revealing what might be classified as a 'defensive epistemology'—that is undertaking 'policy-based evidence making' in the pejorative sense of the phrase (Torriti 2010). Clearly, on the basis of good democratic standards, any policy institution that subscribes to a 'defensive epistemology' is effectively attempting to bypass accountability to citizens and is therefore not a legitimate, defensible position. As such, it is important to identify what a legitimate epistemic position might be for policy institutions with respect to evaluation (i.e. use an inductive approach). In so doing, any failures to deploy stronger tests of policy effectiveness can be identified and new prescriptions for improvement

offered, tailored to the energy policy epistemology. Below, the Green Deal policy (developed by DECC between 2012 and 2015) ex ante and ex post evaluative research and data collection approaches are examined to shed light on what might constitute an 'energy policy epistemology'.

The Green Deal evaluation programme

The Green Deal policy aimed to increase the uptake of insulation and other energy efficiency measures by providing first an official assessment of need (a 'Green Deal assessment') which then identified what measures could be funded by a Green Deal loan. The loan was to be paid back by savings on energy bills generated by the installed measures. Specific suppliers of Green Deal measures were accredited as were the Green Deal assessors. The Energy Company Obligation (ECO) placed an obligation on energy companies to supply energy efficiency measures into low-income homes at no cost to the occupant—while important, little will be made of this element due to space constraints.

The UK's gov.uk website collates analytic outputs centrally making straightforward the discovery of evaluative research and data collection on a policy. There is a single webpage dedicated to the Green Deal and ECO evaluation programme[4] which collates 9 separate projects each of which have published between one and 12 reports. The project names are presented in Table 2 as an indication of the areas of the important questions policy asked in the evaluation activity.

In addition, a dedicated statistics web portal[5] collates all official statistical outputs. A search in the webpage's text box labelled 'Contains', using 'Green Deal' as a search term and 'Policy area' set to 'Energy', returns 79 hits (as at 31 October 2017), each for a monthly statistical return on the deployment of the policy. A wider search for 'Green Deal' on the publications section of gov.uk ('Contains' set to Green Deal, Publication type set to 'Research and Analysis', Policy area set to 'Energy') returns 43 hits, of which 28 are Green Deal related. Fourteen of these are part of the Green Deal and ECO evaluation page, 11 of these associated with Green Deal Assessment research and 4 of which are part of a 'Household tracker' survey. Table 2 summarises these studies with an indication of the study design type, main method classification (quantitative, qualitative or mixed) and an indication of scale if fieldwork is involved.

The first point to make is that there are no RCT-based experimental designs here—the closest RCT in this general policy area is a loft clearance 'behavioural trial' (DECC 2013) linked to the Green Deal insofar as one aspect of the policy was support for loft insulation. This project was carried independently of the main Green Deal evaluation programmes and so is not listed in Table 2. Volumetrically, the greatest emphasis here is on the monthly statistical collection that documents the appearance of Green Deal activity across the UK. What is notable is that these are published publically and badged as 'national statistics' rather than being solely internal monitoring data without any such badging. The 'national statistics' badging is a way of quality assuring government data, in order to ensure they are 'trustworthy' (i.e. valid). It is clear that these data are more than simply monitoring data—they perform a function of demonstrating that the policy is existing and is growing, or, as ultimately happened in this case, tailing off. Ultimately, this data exists as a form of *accountability*—to Parliament, the media, the opposition parties and the general public. Without this data, it would be difficult for the government to show that it has done what it said it would do, and would therefore lose the legitimacy necessary to govern.

The central importance of this data, as determined by the scale of data collection and waves of publication, means the logic outlined above merits further inspection. First, do these statistics provide a useful and valid form of accountability? Or might they hide a level of deadweight in the policy (i.e. many such installations would have happened without the policy) that they are effectively useless? While the data show the amount of Green Deal-branded activities, there is no contextual data showing the number of similar such installations that are not Green Deal-driven making it difficult to know if the Green Deal was adding to or simply 'rebranding' current normal activity that would have happened anyway. Despite this, there are some reasons why these figures alone could provide a reasonable, if sub-optimal, indication of policy success or failure. The first is whether the rate of installation—the numbers per month or year—are increasing or decreasing compared to previous data on related prior policy (as indicated above, this was a critical element of the evaluation of

[4] https://www.gov.uk/government/collections/green-deal-and-eco-evaluation

[5] https://www.gov.uk/government/statistics

Table 2 Evaluative research and data collection related to the Green Deal and ECO policy, showing study design, methods, number of waves and scale of data collection where appropriate

Research/evaluation project title	Design	Methods	Waves	Scale (*n*)
Green Deal assessment research	Survey	Quantitative	3	1500 (500/wave)
Green Deal customer journey surveys	Survey	Quantitative	5	Ca. 4100 (400–900/wave)
Energy Companies Obligation (ECO) customer journey research	Mixed (Survey/interviews)	Quantitative and qualitative	1	571/28
Green Deal household tracker survey	Survey	Quantitative	6	15,000 (3000/wave + 2 1500 small waves)
Research into businesses that were not certified Green Deal suppliers	Mixed (Survey/interviews)	Quantitative and qualitative	1	400/15
Green Deal assessment mystery shopping research	Mystery shopper	Qualitative	1	46
Green Deal pre-assessment customer journey qualitative research	Focus group	Qualitative	1	6 groups
Green Deal provider market report	Interviews	Qualitative	1	39
Evaluation of the Green Deal Communities Private Rented Sector funding	Mixed	Qualitative and quantitative	1	44 interviews, Administrative data covering 23,000 properties
Green Deal and ECO statistics	Administrative data	Quantitative	78	Population statistics of Green Deal installations

policy success). If one assumes that such installations are not going to take place on account of market failure (e.g. split incentives or co-ordination failure), then the likelihood of high deadweight might be seen as a low risk. In that context, what also becomes important is whether the rate of installation indicates sufficient pace of delivery to meet legally binding targets for greenhouse gas emission reductions that had been set in a previous 'Carbon Plan' (DECC 2011a). In addition, impact assessments are published prior to the launch of any new policy in the UK, where the anticipated cost-benefit ratios are identified based on assumptions about rates of installation and the logical value of benefits attributable to those installations (DECC 2011b). As a consequence, should the rate of installations under the Green Deal fall short of the anticipated volume of activity identified by the low end of estimates in the impact assessment then the policy is likely to be killed off. In addition to the deadweight issue above, the other major epistemic challenge is in knowing whether Green Deal (or indeed any equivalent) installation leads to energy demand reductions (and consequent carbon emissions reductions) of the scale assumed in the impact assessments.

The next obvious focus is on surveys: half of the research listed in Table 2 uses a survey design collecting quantitative self-report data. This maintains the focus on quantitative data but adds a subjective element to the representation of the public. Some of these surveys look at the experience of receiving the policy, and some about the perspective of those who have no connection to it. The scale of the surveying (total around 21,500 survey participants) gives rise to a desire to *represent* a population—and to do so in enough detail such that different sub-groups issues and perspectives are taken into account. In the list in Table 2, there are four populations being represented—consumers who have had contact with the Green Deal, consumers who have not had contact, businesses involved in delivering the Green Deal, businesses who are not. This has both a democratic function with regard to enabling different voices to inform the policy, but also a practical function: to improve the policy by addressing potential barriers such as lack of awareness, lack of supply and so on. This practical function is visible in the final obvious feature of the list of studies—the qualitative nature of the designs.

Qualitative data features in six of the studies so is in some sense more common (as a design choice) than the quantitative data. There are three studies that are exclusively qualitative—covering both the consumer and provider side of the market. This reflects a clear privileging for subjectivity in the form of personal perspectives with regard to how and whether the policy is working. In a sense, this perspective can be classed as 'useful subjectivity'. The goal here is clearly not to represent the populations of interest but to gather narrative accounts

of and views on how the policy operates. It is here where a causal mechanism is in part understood, and it is important to note that qualitative inquiry is the preferred mechanism for capturing it, in and alongside significant quantitative data collection.

The pattern of Green Deal research designs, methods and scale elucidates some features of an 'energy policy epistemology'. In democratic states such as the UK and USA, these generate three specific epistemological drivers that promote the use of certain designs. Table 3 sets out the three drivers and the kinds of designs and methods it promotes.

What is absent from Table 3 is a driver that would actively promote (or prevent) an RCT or experimental designs more generally. The likely home for such a driver would be under accountability—to demonstrate to Parliament, the media and the public at large that the government did *actually* cause the observed outcome. Yet the clear indication here is that accountability seems to stop short of that kind of epistemic demand, implying there are other features which need to be invoked if the current set of epistemic drivers is deemed both necessary and sufficient to explain a legitimate lack of interest in RCTs. There are potentially two reasons for this. The first is the role and presence of Chief Scientific Advisers in UK government departments (and associated engineering research teams) potentially leading to a problematic over-reliance on pure physics assumptions that such interventions work in all circumstances, as Vine et al. suggest. In addition to this, two additional inter-related epistemic drivers are proposed below that would actively reduce interest in RCTs: *limited agency* and *negotiated certainty*. Further critical analysis and potentially new empirical research will of course be required to determine whether these concepts carry any explanatory useful power.

Given the particular open nature of the energy systems in democratic states where actors within these systems are not under direct control from the state (in contrast, for example to the health or education sectors in the UK), but are in some sense partners delivering and receiving services with the support and oversight of the state, certain legitimate positions can be held. These comprise the following:

> *Limited agency*—the recognition, especially within the energy system, that policy institutions have limited agency on account of the open nature of the system. This leads to a focus on policy-specific *outputs* (e.g. Green Deal assessments, installations and so on, linked to accountability) rather than outcomes (e.g. energy bill savings, increased thermal comfort, reduced fuel poverty) even though these are guiding goals of policy. Outcomes are affected by a range of external factors which policy institutions in free-market economies are not expected to try and control (indeed it is preferred that

Table 3 Epistemological drivers for an 'energy policy epistemology' and the kinds of designs and methods they promote

Epistemological driver	Why	Designs/methods
Accountability	Being able to demonstrate that the promises of action represented in policy announcements have been undertaken is key to retaining legitimacy and therefore power. The checks and balances built into democratic systems demand a summative demonstration—have you done what you said you would do?	Complete and systematic administrative data collection (census), quantitative, independently quality assured.
Representation	Ensuring an understanding of how a policy affects different groups is critical to retain political legitimacy and credibility. This asks the question, how true is the problem/impact generally and for whom?	Large-scale surveys (in the 1000s); mainly quantitative data; self-report via questionnaire. Likely involve some form of random probability sampling.
Useful subjectivity	Recognition that systems are constituted of sentient actors whose perspectives are both *important* (democratically) and *informative* (pragmatically) regarding the action of the policy. This asks the question, is the policy working well? If not, how can it be improved?	Various forms of qualitative inquiry mainly including interview methods, but also observational methods (e.g. mystery shopper). Often purposively sampled in relation to sub-groups identified via the quantitative survey.

they do not control) at least in domains like energy supply. This limits the degree to which policy institutions must strongly show they directly caused certain outcomes is a relevant or fair question. The concept of limited agency is not about accepting that the world is complex (and therefore unknowable)—RCTs are designed to help manage that complexity via random assignment—it is about power and control, and the way in which energy policymakers in democratic states see their role in shaping society and therefore the way in which certain ways of knowing are privileged.

Negotiated certainty—Given that state actors like policymakers have limited agency in the energy sector, in order to generate causal outcomes, policy actions must be negotiated with stakeholders. This means negotiating future regulatory environments (e.g. the range of conditions under which a householder can install insulation) with business and citizen actors to ensure causal conditions can occur (that is, at the very least that the business community are amenable to supply insulation measures via the proposed programme in the case of the Green Deal). Attempting to unilaterally impose a particular way of doing into a setting where there is widespread acceptance of limited state agency is likely to directly count against policy effectiveness (and is therefore a key external validity threat to any RCT in this area). A good example of this outside the energy sector in the UK was the attempt by the Coalition government in the UK to sell off publicly-owned forest without negotiating the policy with stakeholders.[6] Following widespread criticism, the policy was dropped. This negotiation of certainty is less about imposing strict conditions in a top-down way, but about generating policy impact via bilateral relations and through that negotiation, in effect creating causal effects by agreement.

Both these reasons would count against promoting RCTs in policy research design. The acceptance of limited agency is related to, but different from, the inability of RCTs to effectively deal with threats to external validity (Allcott and Mullainathan 2010). It relates to a political, normative choice about how the state *should* interact with citizens. In the UK energy system which is privatised in delivery and private in consumption, state interference in (for instance) the provision of home insulation is preferred to be relatively limited (compared to the provision of health or justice systems in the UK). As such, any research which fails to take account of the limited agency of the state in this respect goes directly against this approach. Likewise, when policy-making is seen as a means of generating negotiated certainty, then a mode of research such as RCTs which is, in effect, about imposed certainty (insofar as the control of treatment groups and the exact enactment of the treatment itself goes) it is understandable that RCTs are not a preferred way of knowing about causal effects—especially if the act of imposition itself kills off the very causal mechanism intended to be studied.

It is worth noting two important caveats surrounding the presumed existence of these concepts: the first is evidential. Here, no direct empirical data is presented to support the existence of these concepts: they are inferred via a combination of the limitation of the epistemic drivers in Table 3 to fully explain the lack of RCTs, and drawing on the author's personal experience from working in the UK government. Clearly, targeted data collection and analysis to determine whether they are real is needed if these concepts survive initial critical inspection. The second is that the impact of these concepts, if they are important, may be policy-area specific. As implied above, they may arise specifically in democratic states and in particular in domains that are deemed to be delivered not by state actors, such as energy, environment and agriculture and transport. This implies that if they hold as explanatory factors, there ought to be more RCTs in domains with heavy state involvement or control, such as education, health or criminal justice (in the UK).

The 'energy policy epistemology' no doubt has roots in many other scholars' attempts to classify epistemic positions in science. There are clear signs of a critical realist perspective in the privileging a 'useful subjectivity' as a major source for understanding of causal mechanisms and the implicit attempts to understand what works for whom (Pawson and Tilley 1997). Others have created categories that could place this description more widely as part polling democracy, part critical pragmatism (Tapio and Hietanen 2002). Negotiated certainty has echoes of social constructionism (Berger and Luckmann 1991). Either way, it is a long way from the positivist position underpinning RCTs. For now, the

[6] See http://www.independent.co.uk/environment/nature/controversial-plans-to-sell-off-england-s-public-forests-abandoned-by-government-7907605.html Accessed 10 Jan 2018

goal here is simply to elucidate some of these issues to enable a more fruitful dialogue between academics and policy makers.

Identifiable problems within the 'energy policy epistemology'

There are no doubt major epistemic problems associated with the way DECC deployed research designs to inform the evaluation of the Green Deal. Part of this is down to some of the failures to understand causality effectively around outcomes rather than just outputs: what would work to deliver energy savings and improved comfort, rather than just deliver policy at a certain scale in a specific way. This issue is a broader one for what was DECC (now part of BEIS—the department for Business, Energy and Industrial Strategy) and its lack of data describing the wider social context of energy supply and demand. This lack of data around what might be called the outcome space, rather than just the outputs space, is also an issue regarding the ability to capture 'spillovers' as Vine et al. call them. This means there is an accountability gap with regard to the deployment of studies that can effectively capture, with sufficiently strong internal validity, the causal effects of a policy on both the intended and unintended outcomes. Given the strong reasons against deploying RCTs in this context, what might be a better way of strengthening the evaluative capacity of policy organisations given the 'energy policy epistemology'?

The way forward: if not RCTs, then what?

In one sense, we could frame the question thus: what research design approach is likely to bring about the benefits that RCTs are intended to give, while taking seriously the constraints imposed by the legitimate aspects of the 'energy policy epistemology'? In addition, any proposal would need to address outstanding barriers identified above around the issue of timing noted above.

Table 4 describes a systematic, in part quantitative, large-scale data collection exercise that is 'always on' or continuous that can capture natural experiments. This essentially describes a large longitudinal panel design that includes significant subjective self-report facility (so not, for example, simply the collection of smart meter data). This kind of design is even rarer in the policy landscape globally than RCTs (Elam et al. 2014). There are clear benefits to such an approach which the author, with other colleagues have explored for the UK government (Cooper et al. 2014). There are no doubt enormous challenges for developing such an infrastructure but the benefits are likely to be even larger.

Table 4 Identifying the features of a research design that takes advantage of the 'energy policy epistemology' while delivering the benefits of RCTs

Must support (or not undermine)	Research design response
Accountability	A systematic data collection process that reveals true roll out of policy
Representation	Big enough *N* to include significant capture of a range of sub-groups
Useful subjectivity	Uses self-report across a range of stakeholders; supports qualitative inquiry
Limited agency Negotiated certainty	Allows policy delivery variation as a key way of recognising stakeholder agency (not top-down control) and free-market values. Also links to preference for high external validity across varied contexts
Additionally…	
Ability to capture wider outcomes	Be able to link to other data sets
Timeliness	Continuous data collection
High internal validity on causation	Be able to capture natural experiments

Conclusion

Vine et al.'s original paper makes a significant contribution to the debate about how governments should go about determining if a policy is effective. Through promoting the use of RCTs, they raise important questions about what are the drivers for knowledge in this context. The case study of DECC and the Green Deal suggest that it is not necessarily the case that policy organisations share the same perspective on what constitutes good data. This may only be true for the UK, but the claim implied here is that the emergent 'energy policy epistemology' or elements of it are likely visible in other states that share a parliamentary democracy and alignment with a free-market economic ideology. Two important points need to be made:

1. Further research exploring the implicit and explicit drivers for knowledge within policy organisations is needed. For too long, academics from outside these institutions have attempted to infer what happens inside policy organisations regarding the use of evidence. Only limited progress has been made, and more field research is required to determine whether the 'energy policy epistemology' is real, and if so how widely held it is.
2. If the 'energy policy epistemology' is real, then the subsequently proposed remedy to challenges within that framework—a demand for large-scale longitudinal designs—is only one possible response. It appears to be an important one; however, given the way, such a design is able to speak to a variety of challenges in this context. But crucially, the deployment of a large quantitative survey should not preclude the continued use of ethnographic and other qualitative approaches both within and without the policy and academy contexts. This includes the use of RCTs—where they can add valuable insights, they should be deployed.

The hope here is that in setting out a nascent framework for how the academy can think differently about the kinds of research they do and promote, more impact and better policy are the result. To that end, Vine et al. and this author are in complete agreement.

Acknowledgements Although not commonly allowed, I would like to express my deep thanks to the patient and direct referees who enabled me to make this article far superior than it would otherwise have been.

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

References

Allcott, H., & Mullainathan, S. (2010). Behavior and energy policy. *Science, 327*(5970), 1204–1205. https://doi.org/10.1126/science.1180775.

Berger, P. L., & Luckmann, T. (1991). *The social construction of reality: a treatise in the sociology of knowledge, New Ed edition. ed*. Harmondsworth: Penguin.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T., & Moore, L. (2012). Realist randomised controlled trials: a new approach to evaluating complex public health interventions. *Social Science & Medicine, Part Special Issue: Place, migration & health, 75*, 2299–2306. https://doi.org/10.1016/j.socscimed.2012.08.032.

Bothwell, L. E., Greene, J. A., Podolsky, S. H., & Jones, D. S. (2016). Assessing the gold standard—lessons from the history of RCTs. *New England Journal of Medicine, 374*(22), 2175–2181. https://doi.org/10.1056/NEJMms1604593.

Bowen, S., & Zwi, A. B. (2005). Pathways to "evidence-informed" policy and practice: a framework for action. *PLoS Medicine, 2*(7), e166. https://doi.org/10.1371/journal.pmed.0020166.

Cooper, A. C. G., Shipworth, D., & Humphrey, A. (2014). UK Energy Lab: a feasibility study for a longitudinal, nationally representative socio-technical survey of energy-use (Synthesis Report). UCL. http://www.ucl.ac.uk/steapp/docs/lukes-reports/synthesis. Accessed 16 May 2016.

DECC, (2014a). Advice on how to use heating controls: Evaluation of a trial in Newcastle. London, UK: HM Government.

DECC (2014b). Evaluation of the DECC/John Lewis energy labelling trial. London, UK: HM Government.

DECC (2013). Removing the hassle factor associated with loft insulation: Results of a behavioural trial. London, UK: HM Government.

DECC (2012). Final Stage Impact Assessment for the Green Deal and Energy Company Obligation. London, UK: HM Government.

DECC (2011a). The Carbon Plan: Delivering our low carbon future. London, UK: HM Government.

DECC (2011b). Green Deal Impact Assessment. London, UK: HM Government.

Elam, S., Hübner, G. M., Shipworth, D., Shipworth, M., Humphrey, A., & Hamilton, I.G. (2014). UK Energy Lab: feasibility study final report—annex B - Available Data. UCL. http://www.ucl.ac.uk/steapp/docs/lukes-reports/annex-b. Accessed 16 May 2016.

Ettelt, S., & Mays, N. (2015). RCTs: how compatible are they with policy-making? *British Journal of Healthcare Management, 21*(8), 379–382. https://doi.org/10.12968/bjhc.2015.21.8.379.

Ettelt, S., Mays, N., & Allen, P. (2015). Policy experiments: Investigating effectiveness or confirming direction? *Evaluation, 21*(3), 292–307. https://doi.org/10.1177/1356389015590737.

Frederiks, E. R., Stenner, K., Hobman, E. V., & Fischle, M. (2016). Evaluating energy behavior change programs using randomized controlled trials: best practice guidelines for policymakers. *Energy Research & Social Science, 22*, 147–164. https://doi.org/10.1016/j.erss.2016.08.020.

Halpern, D. (2015). Inside the nudge unit: how small changes can make a big difference. WH Allen.

Haynes, L., Service, O., Goldacre, B., Torgerson, D., (2012). Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. Cabinet Office Behavioural Insights Team, London.

Kingdon, J.W., 1984. Agendas, alternatives, and public policies, update edition, with an epilogue on health care, 2 edition. ed. Pearson, Boston.

Marchal, B., Westhorp, G., Wong, G., Van Belle, S., Greenhalgh, T., Kegels, G., & Pawson, R. (2013). Realist RCTs of complex interventions—an oxymoron. *Social Science & Medicine, 94*, 124–128. https://doi.org/10.1016/j.socscimed.2013.06.025.

Oliver, K., Innvar, S., Lorenc, T., Woodman, J., & Thomas, J. (2014). A systematic review of barriers to and facilitators of the use of evidence by policymakers. *BMC Health Services Research, 14*(2). https://doi.org/10.1186/1472-6963-14-2.

Pawson, R., Tilley, N. (1997). Realistic Evaluation. SAGE Publications Ltd, London ; Thousand Oaks, Calif.

Tapio, P., & Hietanen, O. (2002). Epistemology and public policy: using a new typology to analyse the paradigm shift in Finnish transport futures studies. *Futures, 34*(7), 597–620. https://doi.org/10.1016/S0016-3287(02)00003-4.

Torriti, J. (2010). Impact assessment and the liberalization of the EU energy markets: evidence-based policy-making or policy-based evidence-making? *JCMS: Journal of Common Market Studies, 48*(4), 1065–1081. https://doi.org/10.1111/j.1468-5965.2010.02089.x.

Vine, E., Sullivan, M., Lutzenhiser, L., Blumstein, C., & Miller, B. (2014). Experimentation and the evaluation of energy efficiency programs. *Energy Efficiency, 7*(4), 627–640. https://doi.org/10.1007/s12053-013-9244-4.