Evaluation good practice: is 'good enough' better than 'perfect'?

Dr Joanne Wade
Freelance sustainable energy consultant
107 Queens Rd
London SW19 8NR
UK
joanne.wade09@gmail.com

Dr Nick Eyre
Environmental Change Institute, University of Oxford
Oxford University Centre for the Environment
Oxford OX1 3QY
UK
nick.eyre@ouce.ox.ac.uk

Keywords

best practice, domestic energy efficiency, energy efficiency programmes, engineering estimates, randomised control trial (RCT), evaluation methods

Abstract

The 'gold standard' of perfect evaluation practice may in theory be defined as the robust implementation of a Randomised Control Trial (RCT). In reality, the implementation of such an approach has rarely been possible for energy efficiency programmes, and evaluators have delivered studies that are 'good enough', within constraints defined by programme design, evaluation budget and timeframe, and evaluation aims.

Drawing on a systematic review of the peer-reviewed literature on household energy efficiency evaluation undertaken for the UK Energy Research Centre, this paper debates priorities for future evaluation research, based on an analysis of possible gaps in knowledge and evaluation practice. It assesses the benefits and drawbacks of different evaluation methods (including RCT, quasi-experimental methods, and engineering estimates) in terms of cost, complexity and accuracy (in the context of impact evaluation of programmes or policies). It identifies the potential shortcomings of differing methods (e.g. reliance of engineering estimates on deemed savings and the availability of suitable sample sizes for experimental approaches). It sets outs some key gaps in our knowledge about the impacts of energy efficiency programmes, which pose new challenges for evaluation, including assessment of how impacts vary across end-users (rather than just average effects) and assessment of wider market transformation by large scale programmes.

It concludes that there is a role for a range of different evaluation approaches from rigorous RCT under tightly controlled conditions to information collection from major programmes, and that greater efforts are needed to share and debate existing information, via peer review and publication.

Introduction

Energy efficiency programme evaluation in Europe is evolving and a more professional evaluation community is developing (Vine and Thomas, 2012). At the same time, energy efficiency policy in the UK and elsewhere is undergoing a period of substantial change, for example through the introduction of finance mechanisms and weakening of energy company obligations. Multiple policies and programmes have been employed in the past to encourage improvements in household energy efficiency, and many evaluations have been undertaken. However, the accuracy of the approaches used has been questioned by some commentators and theorists and practitioners have differing perspectives about what are appropriate and robust evaluation methods (contrast for example, Frondel and Schmidt 2005 with CPUC 2006).

Policy and programme developers need to have confidence in the information they are getting from evaluations. Therefore it is useful to review what we can learn from existing evaluations, in terms of the benefits and drawbacks of different evaluation methods and hence their usefulness in closing the gaps in our existing knowledge.

This paper focuses on ex-post impact evaluations of policies and programmes to improve the energy efficiency of homes. It does not consider the evaluation of the effects of single energy

efficiency actions (for example, insulating a single house)[1]. It is based on a systematic review of household energy efficiency programme evaluation literature, completed in the summer of 2014. It presents the main evaluation methods in use and discusses the circumstances in which they are most appropriate and the situations in which they are currently most often used.

The paper then reviews what we already know about household energy efficiency programme effects, to identify gaps that need to be addressed. The methods available are then matched to these gaps. Based on this matching, the paper discusses whether a potential 'gold standard' for perfect evaluation – a Randomised Control Trial[2] – is something that we should be aiming to use as often as possible, or whether there are a series of 'good enough' alternatives that can answer our questions as well, if not better.

Method

The findings presented in this paper are based on a systematic review of peer-reviewed programme evaluation literature[3]. This literature was used in two ways: first, to explore the different evaluation methods in use and how these relate to 'good practice'; second, to examine what we know – and hence what we don't know – about the effects of household energy efficiency programmes on energy use.

The evidence used to examine what we know about programme effects was filtered to ensure that only the most robust findings were included. This process involved the project team's expert judgement based on the following series of questions:

- Was the evaluation based on a robust understanding of how the programme was likely to lead to changes in energy use?
- Was the scale and nature of the evaluation appropriate to the programme?
- Was the evaluation method appropriate given the quantity and quality of evaluation data available to the evaluators?
- Did the evaluation acknowledge its own limitations and explore whether these could have had an important effect on the accuracy of the results?

LIMITATIONS

The evidence base was restricted to peer-reviewed papers for a number of practical reasons: the substantial volume of other literature is often not catalogued in databases, and hence is difficult and time consuming to find; the process of peer-review should, in theory at least, help to guarantee a minimum level of quality in the material being used; and literature in languages other than English was not accessible to the authors due to their own limited language capabilities. This does mean that the findings are not fully representative of the state of knowledge. For example, there may be a bias in the topics covered, since peer reviewed papers present original research and hence may often poorly represent evaluations of programmes that are considered well-understood. Similarly, there may be reluctance amongst evaluators or their clients to communicate information about evaluation or programme failures and hence these may not be captured in the literature. However, we are confident that the findings nevertheless provide useful insights. The conclusions to this paper include suggestions of where further work could usefully add to these.

Evaluation methods in use: their benefits and limitations

This paper is concerned with the effect of energy efficiency programmes on energy use; hence it is interested in the extent to which different evaluation methods provide a robust estimate of how a programme changes energy use. The methods considered here are: engineering approaches (simple and enhanced); simple before-after comparison; quasi-experimental approaches (cross-section, difference-in-difference, and with exact matching) and experiments (Randomised Control Trials)[4]. The benefits and limitations of different approaches are linked to two things: how well they can in theory capture the true effect of a programme; and how well they can be used, given the practical realities facing evaluation teams.

EVALUATION THEORY

An ideal ex-post evaluation of a household energy efficiency programme would compare the post-programme energy use of a suitably sized sample of households affected by the programme with what the energy use of the same sample of households would have been if the programme had not happened (the 'counterfactual'). It is usually *relatively* easy for a well-designed evaluation to gather information on the post-programme energy use of participant households. But the counterfactual cannot be measured because it has not actually happened, so the evaluator has to find a way to estimate it. Evaluation theory (see for example, Frondel and Schmidt, 2005; Vreuls, 2005; Vine *et* al, 2012) suggests that, for the estimate to be reasonable, it needs to account for each of the following factors:

- The extent to which exogenous factors (i.e. things other than the programme) are affecting energy use. For example, if household energy use has reduced, is some of this due to energy price increases rather than the effect of the programme?
- Additional programme affects on household energy use (known as participant spillover). For example, someone who takes up a programme offer of a rebate on an efficient refrigerator might then decide to buy an efficient freezer or washing machine because the programme has raised their awareness of the benefits of this, but without any further financial incentive from the programme.

1. For more on the different requirements of measure versus policy/programme impact evaluations see Broc *et al*, 2009.

2. Randomised Control Trials have for many years been seen as the 'gold standard' for evaluation of clinical interventions. Although there are very significant differences between a clinical intervention and an energy efficiency programme, it is useful to consider the extent to which this theoretically accurate method can improve the quality of knowledge about the effects of energy efficiency actions.

3. More detail on the method can be found in the project report, which is available here: http://www.ukerc.ac.uk/programmes/technology-and-policy-assessment/energy-efficiency-evaluation.html

4. These evaluation methods are described in more detail in the project report.

- The extent to which programme-induced energy savings are offset by increased use of the energy service concerned, because it is cheaper, or increased use of other energy services as energy bill savings are spent on other things (known as rebound).
- The extent to which both observable and unobservable differences between households affect not only the way they react to a programme, but also the likelihood of them taking part in the programme (known as self-selection). For example, households that are more willing to make changes to save energy may be more likely to take part in a programme than the average household and also likely to save more energy as a result of participation than the average household.
- The extent to which participating households would have taken programme-supported actions to save energy even without the programme (free-ridership, or lack of additionality).
- The extent to which the programme affects energy use in households that are not considered participants – including through wider, market transformation impacts (non-participant spillover). For example, a household may choose to purchase an efficient appliance as a result of programme marketing or the offer of a rebate, but then not claim the rebate and hence not be considered a programme participant. In some cases, such as education and community programmes, distinguishing between participants and non-participants may not be easy, or even helpful.

EVALUATION PRACTICE

An evaluator must consider not only how well each available method can reflect each of the above factors, he/she must also think about whether or not the method can be used, given the practicalities of an evaluation. Firstly, the method will have to be chosen and defined according to the purpose or objectives of the evaluation. After this, the main limitations on free choice of evaluation method are data quantity and quality; and time and money allocated for the evaluation:

- Evaluation aims can vary significantly within the general aim of understanding the effect of a programme on energy use: a utility may commission an evaluation to meet regulatory requirements and hence primarily be interested in verifying numbers of measures installed; a government may be interested in how levels of energy saving vary between different types of household; a technology manufacturer may wish to understand how a specific product performs in practice. These varying aims are likely to be best served by different evaluation methods. This means that evaluation methods should be chosen first according to the evaluation objectives.
- Data availability can often constrain the choice of evaluation methods. A lack of access to accurate energy data for the period prior to a programme clearly removes the choice of any method comparing actual before and after energy use in participant households; equally, the large samples of well-matched participant and control group households required for a Randomised Control Trial may simply not be available.
- The budget for an evaluation, and the timeframe over which it has to be delivered, can be important in determining the method chosen. In general, the more accurate a method is thought to be in theory, the more data it will require and hence the more costly it will be to deliver. Also, some methods require a greater elapsed time over which energy use is measured and hence may not be suitable when results are required quickly.

BENEFITS AND DRAWBACKS OF THE MAIN EVALUATION METHODS

Table 1 summarises the benefits and drawbacks of each of the main evaluation methods, in terms of their ability to account for the main factors considered important in theory and the extent to which they fit with the practical constraints mentioned above. Definitions of the methods, and a detailed discussion of how this summary has been determined, are included in the project report (see footnote 1). Based on this summary, the table also proposes situations when it is most appropriate to use each method.

A few key points to note from the table:

- Engineering estimates may be amongst the least accurate approaches to evaluation (see, for example, Hamilton *et al*, 2013) since they do not inherently account for as many elements of programme effects as most other approaches. However, they can offer a very cost-effective approach where the effects of measures have been previously evaluated rigorously, are well understood, and hence can be incorporated through the use of correction factors.
- Randomised Control Trials clearly address more of the different aspects of programme effect than other methods. However, they require very tightly controlled experimental conditions, and large datasets. Hence they are expensive and not appropriate for programmes where tightly controlled conditions are inconsistent with the programme design.
- Non-participant spillover is not addressed by any of the main methods used. This may lead to systematic underestimation of programme effects and is a particular concern where non participants are used as a comparator group (i.e. in quasi-experimental approaches) and where such spillover is an aim of the programme (i.e. market transformation).

The evidence base

Papers reporting programme evaluations were found across a broad range of conferences and publications. The IEPEC, IEPPEC, eceee and ACEEE conferences were the richest sources of material, containing over half of all the papers found, but the remaining papers were spread across 20 different publications in energy, building science, energy economics and environmental science/geography.

As should be expected from a peer-reviewed set of literature, the majority of papers reviewed seemed to be of a high quality. However, two points should be noted.

First, the papers gave very little information about the context within which the evaluated programmes had been implemented or any detail about the socio-demographics of the households targeted by the programme: they tell us something about what happened, but only a limited amount about why

Table 1. Summary comparison of methods.

Method	Issues in defining the counterfactual						Key benefits	Key drawbacks	When to use
	Exogenous influences	Participant spillover	Rebound	Self-selection bias	Free-ridership	Non-participant spillover			
Simple engineering	?	?	?	x	x	x	Very few data to collect; cheap	Inaccurate	As cross-check when no better data available
Enhanced engineering	?	✓	?	x	✓	x	Relatively few data to collect; relatively cheap	Potentially less accurate than quasi-experimental approaches	As cross-check; when measures well understood; when interaction between measures is of interest
Before-after	x	✓	✓	✓	x	x	Requires participant group only	Does not account for exogenous influences	When there is unlikely to be much variation in exogenous influences; when a comparator group cannot be found
Quasi-experimental (cross-section)	?	✓	✓	x	?	x	Does not require 'before' data	Needs data from comparison group; non-participant spillover can cause inaccuracies	When 'before' data are not available and when non-participant spillover is not likely to be large
Quasi-experimental (difference-in-differences)	?	✓	✓	x	?	x	Does account for some of the effect of exogenous influences	Increased data requirements; non-participant spillover can cause inaccuracies	Where there is good availability of data for participants and non-participants and when non-participant spillover is not likely to be large
Quasi-experimental with exact matching	✓	✓	✓	✓	?	x	Has the potential to accurately account for self-selection bias	Data requirements may make impractical; non participant spillover can cause inaccuracies	When large datasets are available; where non-participant spillover is not a major issue
Experiments (Randomised Control Trials)	✓	✓	✓	✓	✓	x	Has the potential to provide the most accurate estimate of programme impact on participant households	Can only be used where implementation conditions can be tightly controlled	For pilots of new interventions where there are unlikely to be non-participant spillover effects

it happened. Hence it is impossible to generalise the results of evaluations into likely impacts of programme implementation in other locations.

Second, the evaluation literature scored significantly less well against the final assessment question (acknowledgement and discussion of evaluation limits) than against the others (see 'method,' above). Only a little over half the papers reviewed fully recognised the limits of the evaluation method used, and very few discussed the implication of these for the robustness of the results.

How different programmes types are evaluated

For convenience, household energy efficiency programmes are categorised here into minimum efficiency standards for buildings; energy labelling of buildings; appliance market transformation activities; investment and refurbishment programmes; innovative finance mechanisms; information and advice; smart metering and billing feedback[5]; and community-led energy action. Clearly there is some overlap between these categories and many programmes will include more than one of these types of action (for example, a refurbishment programme may well include innovative finance and information and advice), but in most cases, programme implementers and/or evaluators seem to categorise their programmes within one of these headings and hence they are a useful way to group the information available.

MINIMUM EFFICIENCY STANDARDS FOR BUILDINGS

The effects of minimum efficiency standards for buildings have been evaluated using a number of different approaches. The IEA (Saussay *et al*, 2012) and Deason and Hobbs (2012) compare the evolution of space heating energy consumption or energy efficiency across different territories with differing regimes of minimum standards, whilst Kjaerbye *et al* (2011) and Rogan and O Gallachoir (2011) take more bottom-up views of differences in energy use data in homes, in Denmark and Ireland respectively, built to different sets of minimum standards. These are all essentially quasi-experimental approaches, whereas Tiedemann (2012) uses an enhanced engineering approach to estimate the effect of a building code in British Columbia.

All the studies suggest that minimum efficiency standards do lead to reduced energy use, but as a group they tell us little more than this. The bottom-up approaches suggest that minimum standards do not in practice reduce energy use by as much as pre-implementation engineering estimates would suggest, whilst the top-down study that makes a similar comparison concludes that the overall effect of the standards is greater than predicted by engineering estimates.

ENERGY LABELLING OF BUILDINGS

There are very few impact evaluations of building energy labelling in the literature: too few to inform any conclusions about effectiveness. The studies that are reported use cross-section comparison of energy use in carefully matched labelled and unlabelled homes (Kjaerbye, 2009), or surveys of self-reported measures installed by self-selecting householders combined with engineering estimates of measure effectiveness (Herppich, 2011).

APPLIANCE MARKET TRANSFORMATION ACTIVITIES

Programmes in this area offer a clear example of the inadequacies of traditional energy efficiency programme impact evaluation methods. As labelling or minimum efficiency standards affect entire markets, constructing a counterfactual that uses non-participants as a comparison group is impossible. Equally, simple comparison with the existing stock of technologies is unlikely to offer an accurate estimate since autonomous rates of technological change for these appliances are relatively high and hence the energy efficiency of new appliances is unlikely to be similar to that of the existing stock, even in the absence of policy action.

Some of the main evaluations of EU energy labelling and standards (for example, Bertoldi *et al*, 2001) restrict themselves to demonstrating that the introduction of labels and standards coincides with significant increases in the energy efficiency of appliances sold. They do not attempt to separate the effect of these programmes from other influences on efficiency. However, there is a small number of studies that do attempt to construct a counterfactual: Meyers *et al* (2003) combine historical trends and expert judgement to look at the rate of technological change in the US in the absence of standards, whilst Lane *et al* (2007) use market trends and stakeholder interviews to estimate counterfactuals for refrigeration appliance market transformation in the UK and Australia.

INVESTMENT AND REFURBISHMENT PROGRAMMES

There is a substantial amount of information available about the effects of investment and refurbishment programmes, both for low income programmes and those more broadly targeted. However, the majority of this information is linked to utility programmes in the US and UK (where these programmes have dominated household energy efficiency programme activity) and has not been subject to external peer review. The evaluation of these programmes is generally driven by regulatory requirements, and hence follows set protocols (for example, CPUC, 2006). These specify methods to be used (ranging from using deemed savings values based on previous experience and engineering estimates through to various forms of quasi-experimental billing analysis) and a number of correction factors that can be used to account for free-ridership or rebound.

Bottom-up studies in the peer-reviewed literature include examples of cross-section comparison (Bundgaard *et al*, 2013), engineering assessments adjusted based on small samples of observed savings (Rosenow and Galvin, 2011), and difference-in-difference comparisons based on billing analysis (Scheer and Clancy, 2011). The evidence from these evaluations is consistent in showing two things: firstly that programmes do lead to energy savings, and secondly that these savings are significantly lower than would be expected from simple engineering estimates. However, there is little agreement about the extent of the difference between the estimates produced by different methods.

In addition, recent literature has reported on an alternative approach to assessing the effects of this type of programme.

5. 'Information and advice' and 'billing feedback' are treated as separate categories because the former has tended not to start from the actual energy use patterns of the household whereas the latter does, and because billing feedback has been evaluated in a very different way to other forms of information provision.

A number of authors (for example, Horowitz, 2007) propose looking at effects of portfolios of energy efficiency programmes[6]. The studies use economy-level data on energy use and longitudinal or cross-sectoral comparisons of times or territories with different levels of investment in energy efficiency. These studies produce widely varying results, with some suggesting that the net effect of programmes is much lower than individual programme evaluations would suggest whilst others suggest that they are much higher.

INNOVATIVE FINANCE

There are few reported evaluations of innovative finance mechanisms. Those that do exist seem to use similar evaluation methods to those used for more traditional investment programmes, and hence may also overestimate the effect on energy use unless appropriate correction factors have been applied.

INFORMATION AND ADVICE

There is very little quantification of the effects of information and advice in the peer-reviewed literature. The evidence that is available is either based on very small sample sizes or relies heavily on self-reported effects from participant surveys. This reliance on self-reporting reduces confidence in the results but, in situations where the advice and information is offered widely (so no comparison group is available) and where other mechanisms are also being used to affect energy use (so before-after comparisons will not offer an accurate estimate), this may be the only method available to the evaluator.

SMART METERING AND BILLING FEEDBACK

Billing feedback has received more attention in recent peer-reviewed literature than any other programme type. The coincidence of large-scale implementation of the approach with the availability of smart meter data has enabled the use of experimental approaches to study its effects. Amongst others, Allcott (2011), Agnew *et al* (2012) and Agnew and Gaffney (2013) report on a relatively large number of studies of programmes in the US, all conducted using Randomised Control Trials. A similar evaluation of a large scale pilot in Europe is reported on by Pyrko (2013).

These evaluations present a reasonably consistent picture of programmes that, on average, result in a small reduction in electricity demand. Specific studies also report on the short-term persistence of the savings and on the large variability in effects across households. There is very little information on the reasons for variation between households or on the relative effectiveness of different feedback methods given in these programme impact evaluation papers.

COMMUNITY-LED ENERGY ACTION

There is very little literature quantifying the overall effect on energy use of community-led energy action. This is perhaps not surprising since the scale of activity in an individual project is often too small to allow robust evaluation using most of the methods defined here. Community programmes also implicitly aim to have high spillover and therefore tend to have imprecise definitions of participants, leading to difficulties for conventional bottom-up evaluation techniques.

Gaps in our understanding of energy efficiency programme effects[7]

We know quite a lot about the effects of programmes to stimulate investment in energy efficiency technologies, whether through minimum standards or through subsidies. However, even for these types of programme, there remain gaps in our understanding.

Minimum efficiency standards for buildings appear, from bottom-up studies of the buildings affected, to reduce energy use by less than ex-ante engineering estimates would suggest. Evaluation of investment and refurbishment programmes, using a range of the methods defined above, supports the idea that simple engineering calculations overestimate the energy savings that will be achieved. These findings are also supported by Sunikka-Blank and Galvin (2012) in their research on the actual performance of homes with energy certificates. Furthermore, real differences between engineering calculation results and actual energy use are found in individual buildings that are closely monitored, which can be explained by a combination of installation issues and user behaviour, and this understanding is reflected in correction factors that are routinely applied to calculations of the effect of standards and investment programmes (Danskin, 2014). However, the various studies produce differing estimates of the extent of overestimation, and it is likely that different correction factors are applied by different programme implementers and evaluators. From the information given in the literature, it is not possible to understand the cause of the differences in estimates: the differing assumptions and methodologies behind each study may be valid and the programmes reported may differ in the proportion of the theoretical savings they deliver. Alternatively, one study may give a more accurate estimate of programme effectiveness than another.

Also, top-down assessments of minimum efficiency standards tell a different story, suggesting that engineering approaches may underestimate the overall net effect of the standards. The available evidence does not provide any insight into what may be happening that could explain this.

We do not yet have a clear understanding of the effect of energy labelling of buildings. Labelling and minimum efficiency standards for appliances have also proved difficult to evaluate quantitatively, because there is no observable counterfactual for these economy-wide programmes. However, in this case the use of expert views and historical trends has offered some quantification: in the US, Meyers *et al* (2003) estimate that a range of appliance efficiency standards have reduced household primary energy demand by around 8–9 % whilst Lane *et al* (2007) estimate that standards and labelling of domestic refrigeration appliances have reduced household electricity demand

6. This type of study can be used to capture the overall effect of a whole range of energy efficiency policies and programmes implemented in a given territory. However, the papers found in this review generally dealt with areas where policy and programme activity was dominated by investment and refurbishment programmes, and hence we deal with them here.

7. We would like to remind the reader at this point that these findings are based on the knowledge reported in the peer-reviewed literature: it may well be possible to fill some of the gaps identified from the wealth of evaluation findings that have not been discussed within this literature.

in the UK by around 2 %. Here again, the reasons for the differences between these estimates cannot be determined from the information given in the literature studied.

Evaluation good practice guidelines (for example, Vreuls, 2005) recommend the use of multiple evaluation methods to produce a number of estimates for the effects of a programme: similar results from different methods may increase confidence in the accuracy of these results, whilst differing results will highlight the need for further work. The use of top-down assessment of energy efficiency investment project portfolios is an interesting example of this. As mentioned above, there are significant differences between the results of these top-down evaluations and the results of bottom-up individual programme assessments. And there are also significant differences between the different top-down studies. Work on indirect rebound (for example Sorrell, 2007) offers support to the idea that net economy-wide effects may be significantly lower as a result of this effect. However, as Vine (2013) points out, as an increasing proportion of the population is in one way or another affected by energy efficiency programme activities, non-participant spillover from individual programmes is likely to increase, and this will increase net economy-wide effects. As yet, we do not have a good enough understanding of these effects to conclude whether any of the top-down studies offers a more accurate estimate than more traditional individual programme evaluations.

We have recently focused a lot of evaluation attention on billing feedback, with mixed results so far in terms of our understanding of the effects of this type of programme.

Experimental evaluations of billing feedback programmes may have given us a clear understanding of their overall, average effects (a reduction in electricity use of 1–4 %) and facilitated their acceptance as an element of utility energy efficiency activity. However, the difficulty of ensuring that such experiments are robustly implemented should not be underestimated (Darby *et al*, 2011) and reports in the peer-reviewed literature tend to focus on evaluation methods and results, rather than on the practicalities of the experiments themselves, so it is not possible to comment on their quality. The lack of information on why effects vary between programmes and between households may in part be addressed in the process evaluation literature, which was outside the scope of the present study, but there may be a need for more work on this.

Innovative finance, information and advice, and community-led energy efficiency action are all programme areas where significantly more evaluation work is needed for us to understand their effects.

Innovative finance (for example the Property Assessed Clean Energy programmes in the US or the KfW CO_2 Building Rehabilitation Programme in Germany) is still a relatively new type of programme in comparison with many others, and hence it is not surprising that, as yet, we know little about its effectiveness.

It appears that we actually know very little about the effect of information and advice programmes on energy use. This may well be because, in many cases, their impacts are recorded as part of the effect of the programmes that they support (for example, investment programmes). Separating out the effects of different mechanisms within a programme is not something that existing evaluations focus on. Moreover, in some cases, theory and detailed empirical evidence indicate that both investment and advice are required to achieve energy savings effectively, and therefore a combined programme is appropriate and the policy relevant evaluation is of the combined effect.

As stated above, the scale of individual community-led energy activities is often very small. As well as limiting the range of evaluation methods that can be used, this also limits the budget available for evaluation. The lack of quantified outcomes in the peer-reviewed literature may reflect either a lack of robustness in the results of evaluations or simply that the evaluators of the programmes to date have not been interested in sharing the results with the academic community (as may well be the case for other programmes also), rather than an absence of evaluations.

Closing the gaps: the evaluation methods needed

There are two types of gap in our knowledge about household energy efficiency programme effects that require attention from the evaluation community: uncertainties about the effects of individual programmes, and uncertainties about the wider effects of household energy efficiency programmes in general.

EFFECTS OF INDIVIDUAL PROGRAMMES

The key outstanding question specific to large-scale investment programmes and minimum efficiency standards seems to be whether or not the correction factors that are currently applied to engineering estimates, based on the results of earlier evaluations, are 'good enough' to produce reasonable estimates of programme effects. There is a huge body of evaluation literature for investment programmes that is in the grey literature: this may to some degree answer this question. Equally, there is a need for more top-down evaluations of multiple programmes, to further explore the contribution that this approach can make, and further challenge more traditional evaluation methods.

It is difficult to see what alternative there is to use of market trends and expert opinion in the formulation of counterfactuals for appliance market transformation programmes. Hence, the focus here should perhaps be on more studies of different programmes by different researchers, each of whom will have their own views on how best to interpret market trends and how to gather an unbiased expert view. As the number of studies increases, any consistency in results should increase confidence, whilst any significant differences will highlight where more thought is needed.

We probably know enough about the likely average effects of billing feedback programmes. We need to focus now on understanding the variability of outcomes between households and between programmes, in order to improve design of good practice, and also understand the persistence of changes in home energy use over time. This change in focus will also require smaller scale, more focused experiments, perhaps different analysis of the experimental data already collected, and more longitudinal survey work with households. Hypotheses about how programmes are affecting energy use will be needed as part of evaluation design here.

The opportunity to explore the effects of innovative finance mechanisms experimentally has probably passed, since as their application becomes more widespread, comparison group definition will become difficult. Where awareness of generally available financing remains low (for example, some would argue that this is the case for the UK's Green Deal), there may be

scope for some experimental investigation. However, guaranteeing that selected comparison group households have in no way been affected will be difficult. It may be more interesting – and tell us more about programme effectiveness – to look more closely at, for example, the types of household making use of the finance on offer.

Understanding more about the effects of information and advice may well require definition of good methods to separate out the effects of different mechanisms within one programme. Vine (2013) suggests the use of hypotheses of how different programmes affect energy use to allocate portions of overall changes in energy efficiency to each of multiple programmes acting at the same time. This approach may be equally useful for allocating the overall effects of a single programme to the mechanisms within it. However, where mechanisms and/or programmes interact strongly, so that the combined effects are very different from being simply additive, any allocation can be misleading.

Small-scale community-led programmes are another area where the grey literature may answer some of the outstanding questions. However, the level of funding available for evaluation of these schemes and methodological problems inherent in programme design may mean that the results are not particularly robust. With increasing roll-out of smart metering and widespread use of energy labelling for buildings offering new datasets to evaluators (see, for example, DECC, 2013), and increasing data-processing power, there may be new opportunities to apply more macro-level approaches to explore the effects of small-scale programme, for example by comparing the evolution of energy use in areas with high levels of community energy action to that in areas with little or no community-led action.

WIDER EFFECTS OF HOUSEHOLD ENERGY EFFICIENCY PROGRAMMES

The wider effects of household energy efficiency programmes can be split into two elements: indirect rebound affecting energy use outside the home and non-participant spillover.

As Table 1 summarises, some of the usual evaluation methods do take into account the concept of rebound. However, the aspects of rebound captured are the direct effects on energy use within participant households, due to lower cost energy services and/or the re-spending of saved money. The methods do not consider indirect rebound, i.e. the other consequential effects on energy use outside these homes due to wider changes in the economy, such as impacts on economic growth, which may be more significant in reducing the economy level impact of energy efficiency programmes but are much less well understood (Sorrell, 2007).

None of the methods summarised in the table takes into account non-participant spillover. In many cases, including those most amenable to experimental methods of evaluation, this may be a small effect. Randomised control trials are designed specifically to avoid such effects, and some energy efficiency programmes, such as trials of new technology and small scale billing feedback programmes, may have a very low risk of such spillover and therefore are appropriately evaluated in this way. However, this is not generally the case for energy efficiency programmes.

- Appliance market transformation programmes seek to influence the price and availability of energy efficient devices outside the group of direct participants.
- Education and marketing campaigns (especially new viral methods) seek to develop knowledge and/or engagement across wide populations rather than define specific target audiences.
- Community based approaches seek to alter energy using practices socially rather than in specific households.

In all these cases, part of the goals of the programme are to influence actions by people who are considered to be 'non-participants' by conventional experimental evaluation techniques. Experimental evaluation techniques are therefore of more limited value for these programmes.

The priority for improving evaluation is probably to better understand the magnitude of the effects of these two opposing mechanisms: without this understanding, assessing the robustness of macro-level assessments of wider effects of programmes is not going to be possible. This is likely to require greater testing of alternative hypotheses about the mechanisms through which energy efficiency programmes have wider impacts, using large energy use datasets and expert opinion on counterfactuals.

Conclusion

Evaluation theory demonstrates that experimental approaches (Randomised Control Trials) suffer fewer deficiencies, and therefore should produce more accurate estimates of programme effects on participants, than other methods for individual programme evaluation. Therefore, we could simply conclude that we need to do a lot more experiments. However, even the highest quality experimental techniques cannot (and do not attempt to) account for non-participant effects that may be key outcomes of some energy efficiency programmes. In addition they do not attempt to evaluate why people participate or how programmes might be improved. A more pluralistic approach to evaluation than just RCTs is therefore justified.

Our review of the gaps in current knowledge and the work needed to close these has demonstrated that the priorities for evaluation work are:

- To bring information from grey literature into a common, well-understood and debated knowledge base. This will require multi-disciplinary work to ensure that the quality of the literature is fairly assessed and multi-national work to ensure that information from literature in any language can be included;
- More of the same types of evaluation of some programme types, but carried out by a broad range of researchers, using different datasets and with advice from different experts, to increase confidence in the results;
- Use of large datasets and macro-approaches, both to explore the wider effects of multiple programmes and also to offer an alternative view on the effects of individual programmes.

Where does the 'gold standard' of Randomised Control Trials fit into this picture? Obviously, it is one of the methods that could and should be used more. However, as we have discussed above, it is both difficult and expensive, so the opportunities for its use are likely to be limited. It may also not even be the appropriate method for programmes where there is no ade-

quate or meaningful control group, such as market transformation, education, marketing and community programmes. It is not valid to assume that these are, in some way, inferior programme types, simply because they are not amenable to evaluation by techniques developed for the assessment of individualised medical interventions. If there is a case on grounds of potential effectiveness for these types of programme, where social interaction, e.g. through community-led programmes or via social media, is part of the process of change, it is important to develop effective evaluation approaches for them.

We believe this is increasingly possible. The principles of rigour that are embodied in a Randomised Control Trial should be considered in any evaluation, rather than its detailed processes adopted out of context. Many quasi-experimental approaches come close to meeting the accuracy standard of an experiment, and these can increasingly be employed as access to ever more detailed data on energy use becomes possible. And the gap between simpler methods and experimental estimates can be closed with intelligent use of correction factors, if we develop a better understanding of the magnitude of some of the key effects involved.

In any event, what is critical is a theory based evaluation, i.e. the development and testing of hypotheses about how a programme affects energy use (e.g. through improved technical performance, better individual knowledge, different market structures and/or changes in social practices), then, as far as is possible, all the elements of a programme's effect on energy use should be accounted for. The limitations of the approach taken will also be obvious, and their potential effect on the result of the evaluation can be sensibly discussed.

In this way, a range and combination of evaluation methods will be able to produce results that are 'good enough'. A Randomised Control Trial can give a 'perfect' answer to a question that excludes many of the market and social processes that energy efficiency programmes now seek to engage. A 'good enough' answer to the appropriate question is good enough, and probably better than a 'perfect' answer to the wrong question.

References

Agnew K and Gaffney K, 2013, What do we know about comparative energy usage feedback reports for residential customers? *European Council for an Energy Efficient Economy Summer Study.*

Agnew K, Rosenberg M, Tannenbaum B and Wilhelm B, 2012, Home energy report forms: power from the people, *American Council for an Energy Efficient Economy Summer Study.*

Allcott H, 2011, Social norms and energy conservation, *Journal of Public Economics*, 95, 1082–1095.

Broc, J-S, Thomas S, Adnot J, Bourges B, and Vreuls H, 2009, *The developing process for harmonised bottom-up evaluation methods of energy savings*, Deliverable D4, EMEEES project.

Bertoldi P, Waide P and Lebot B, 2001, Assessing the market transformation for domestic appliances resulting from European Union policies, *European Council for an Energy Efficient Economy Summer Study.*

Bundgaard SS, Togeby M, Dyhr-Mikkelsen K, Sommer T, Kjaerbye VH and Larsen AE, 2013, Spending to Save: evaluating the energy efficiency obligation in Denmark, *European Council for an Energy Efficient Economy Summer Study.*

CPUC, 2006, California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals. State of California Public Utilities Commission.

Danskin H, 2014, personal communication with Hunter Danskin, Head of Technical Energy Analysis, UK Department of Energy and Climate Change.

Darby S, Anderson W and White V, 2011, Large-scale testing of new technology: some lessons from the UK smart metering and feedback trials, *European Council for an Energy Efficient Economy Summer Study.*

Deason J and Hobbs A, 2012, Codes to cleaner buildings: effectiveness of US building energy codes, *International Energy Program Evaluation Conference*, Rome.

DECC, 2013, National Energy Efficiency Data-Framework: part II – impact of energy efficiency measures in homes, London: Department of Energy and Climate Change.

Frondel M and Schmidt CM, 2005, Evaluating environmental programs: the perspective of modern evaluation research. *Ecological Economics*, 55, 515–526.

Hamilton, I.G., Steadman, P.J., Bruhns, H., Summerfield, A.J., Lowe, R., 2013. Energy efficiency in the British housing stock: Energy demand and the Homes Energy Efficiency Database. *Energy Policy* 60, 462–480.

Herppich W, 2011, Smart information ignites significant energy savings – evaluation of a large efficiency program: lessons learnt from the utilities perspective, *European Council for an Energy Efficient Economy Summer Study.*

Horowitz MJ, 2007, Changes in electricity demand in the United States from the 1970s to 2003, *The Energy Journal*, 28, 93–119.

Kjaerbye VH, 2009, Does energy labelling in residential housing cause energy savings? *European Council for an Energy Efficient Economy Summer Study.*

Kjaerbye VH, Larsen A and Togeby M, 2011, Do changes in regulatory requirements for energy efficiency in buildings result in the expected energy savings? *European Council for an Energy Efficient Summer Study.*

Lane K, Harrington L and Ryan P, 2007, Evaluating the impact of energy labelling and MEPS – a retrospective look at the case of refrigerators in the UK and Australia, *European Council for an Energy Efficient Summer Study.*

Meyers S, McMahon JE, McNeil M and Liu X, 2003, Impacts of US Federal energy efficiency standards for residential appliances, *Energy*, 28, 755–767.

Pyrko J, 2013, Energy saving targets – tested in households in the Swedish largest electricity saving experiment, *European Council for an Energy Efficient Economy Summer Study.*

Rogan F and O Gallachoir B, 2011, Ex-post evaluation of a residential energy efficiency policy measure using empirical data, *European Council for an Energy Efficient Economy Summer Study.*

Rosenow J and Galvin R, 2013, Evaluating the evaluations: evidence from energy efficiency programmes in Germany and the UK, *Energy and Buildings*, 62, 450–458.

Saussay A, Saheb Y and Quirion P, 2012, The impact of building energy codes on the energy efficiency of residential

space heating in European Countries – a stochastic frontier approach, *International Energy Program Evaluation Conference*, Rome.

Scheer J and Clancy M, 2011, Quantification of energy savings from Ireland's Home Energy Saving Scheme: an ex-post billing analysis, *International Energy Program Evaluation Conference*, Boston.

Sorrell S, 2007, The rebound effect: an assessment of the evidence for economy-wide energy savings from improved energy efficiency, *Technology and Policy Assessment reports*, London: UK Energy Research Centre.

Sunikka-Blank M and Galvin R, 2012, Introducing the prebound effect: the gap between performance and actual energy consumption, *Building Research and Information*, 40, 260–273.

Tiedemann K, 2012, Coding Conservation: does a residential energy code significantly reduce energy and natural gas use? *International Energy Program Evaluation Conference,* Rome.

Vine E, 2013, Transforming the energy efficiency market in California: key findings, lessons learned and future directions from California's market effects studies, *Energy Policy*, 59, 702–709.

Vine E, Hall N, Keating KM, Kushler M and Prahl R, 2012, Emerging issues in the evaluation of energy-efficiency programs: the US experience. *Energy Efficiency*, 5, 5–17.

Vine E and Thomas S, 2012, Introduction. *Energy Efficiency*, 5, 3–4.

Vreuls H, 2005, *Evaluating energy efficiency policy measures and DSM programmes. Volume 1: evaluation guidebook*. International Energy Agency Implementing Agreement on Demand-Side Management Technologies and Programmes.

Acknowledgements

This work was supported by the UK Energy Research Centre supported by the UK Natural Environment Research Council under grant number NE/G007748/1. Eyre also acknowledges generous support for his research fellowship from the Frank Jackson Foundation. We gratefully acknowledge the literature review work undertaken by the project's research assistant, Victoria Bignet. We would like to thank Jamie Spiers and his colleagues in the UK Energy Research Centre's Technology and Policy Assessment group for their advice on the process of systematic review on which this analysis is based. We would also like to thank the original project steering group and peer reviewers for their valuable feedback as the project progressed.